Première analyse des pluies extrêmes dans la région Cévennes-Vivarais

Caroline Bernard-Michel, Laurent Gardes, Stéphane Girard, Gilles Molinié

▶ **To cite this version:**

Caroline Bernard-Michel, Laurent Gardes, Stéphane Girard, Gilles Molinié. Première analyse des pluies extrêmes dans la région Cévennes-Vivarais. 2008. hal-00385415

HAL Id: hal-00385415
https://hal.science/hal-00385415

Preprint submitted on 19 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Première analyse des pluies extrêmes dans la région Cévennes-Vivarais

C. Bernard-Michel — L. Gardes — S. Girard — G. Molinié

N° ????

Octobre 2008

Thème COG

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--????--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

Première analyse des pluies extrêmes dans la région Cévennes-Vivarais

C. Bernard-Michel*, L. Gardes *, S. Girard *, G. Molinié †

Thème COG — Systèmes cognitifs
Équipes-Projets MISTIS

Rapport de recherche n° ???? — Octobre 2008 — 96 pages

Résumé : Ce rapport présente l'estimation des périodes et niveaux de retour des pluies ponctuelles en Cévennes-Vivarais. On montre que la loi de Pareto généralisée (GPD) avec un paramètre de forme positif est la mieux adaptée pour modéliser les valeurs les plus intenses des pluies Cévenoles. Il apparaît ainsi que la distribution de ce type de pluies est à queue lourde, ce qui se traduit théoriquement par une appartenance au domaine d'attraction de Fréchet.
Les estimateurs de maximum de vraisemblance et de Hill sont utilisés pour calculer les paramètres de la loi GPD à chaque station pluviométrique et ainsi déterminer des périodes de retour ponctuelles. Ensuite, les cartes de période de retour pour les pluies horaires et journalières sont obtenues par krigeage des périodes de retour ponctuelles. Ces cartes montrent que les temps de retour sont fortement liés au relief et que ce lien est radicalement différent selon le pas de temps étudié.
Dans la deuxième partie de ce rapport, on montre que l'hypothèse de stationnarité des séries de pluies, sous-jacente à l'utilisation de la loi GPD, est erronée. Le développement d'un modèle prenant en compte la non stationnarité temporelle des pluies est donc recommandé. Nous proposons un modèle basé sur le découpage de la série temporelle en saisons homogènes. Le découpage est réalisé par une approche non paramétrique basée sur la statistique du test de Kruskal-Wallis. Une fois le découpage optimal déterminé, les valeurs les plus intenses de pluies sont modélisées par un mélange de loi GPDs dont chaque composante correspondra à une saison.

Mots-clés : temps de retour, niveau de retour, Fréchet, Gumbel, variogramme, krigeage, pluie, théorie des valeurs extrêmes, géostatistique.

* MISTIS - INRIA Rhône-Alpes
† Laboratoire d'étude des transferts en hydrologie et environnement

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Extreme rainfall analysis in the Cévennes-Vivarais region

Abstract: In this report, return levels and return periods for rainfall are determined for each daily and hourly raingauge of the Cévennes-Vivarais region. It is shown that the generalized Pareto distribution with a positive shape parameter is the most appropriate to model extreme rainfall values in the region. Theoretically, this means that extreme rainfall distibution belongs to the Fréchet family. The parameters of the GPD are estimated by the maximum likelihood method or by the Hill estimator for each raingauge. Return levels and periods are then easily deduced and extended spatially by kriging. Results show that return levels and periods are strongly correlated to landscape with different conclusions according to the time resolution.
However, the use of GPD density relies on an asumption of stationarity for rainfall series that is not correct. We then propose in the second part of the report to look for a model that would take into account non-stationarity. In that model, time series are cut into homogeneous seasons. They are chosen according to a non parametric method based on the Kruskal-Wallis statistic. Separate GPD models are constructed for each season and then aggregated into a mixture model allowing the estimation of return periods.

Key-words: return period, return level, Fréchet, Gumbel, variogram, kriging, rainfall, extreme value theory, geostatistics.

Table des matières

Chapitre 1

Problématique

En raison de la spécificité géographique et climatique du bassin méditerranéen, le sud-est de la France est régulièrement soumis à des précipitations intenses. Elles donnent lieu à des crues violentes et rapides, qualifiées de crues-éclair, dont les répercussions socio-économiques peuvent être importantes. Citons par exemple les épisodes de crues torrentielles qui ont noyé Nîmes en 1988, Vaison la Romaine en 1992 ou encore la Vallée de l'Aude en 1999. Les perturbations méditerranéennes sévissent particulièrement en automne selon un schéma appelé Episode Cévenol (car c'est en général le massif des Cévennes qui reçoit les plus grosses précipitations). Le schéma général d'un épisode cévenol est le suivant : de grandes masses d'air humide provenant de la mer Méditerranée rencontrent dans leur chemin les montagnes cévenoles, plus froides. Il en résulte de grands phénomènes de condensation qui se transforment en pluies torrentielles sur la région cévenole et ses alentours, notamment le Gard ou l'Hérault. Les principaux départements affectés par ces pluies sont ceux ayant une partie de leur territoire dans les Cévennes : l'Ardèche, le Gard, l'Hérault et la Lozère. Autour, l'Aude subit un phénomène proche au pied de la Montagne Noire. Les Bouches-du-Rhône et le Vaucluse sont affectés indirectement lorsque le Rhône déborde de son lit vers l'est sous l'effet du débit augmenté de ses affluents de sa rive gauche. D'autres événements peuvent affecter tous les départements méridionaux mais on parlera alors plutôt d'épisode méditerranéen.

Comprendre ces évènements extrêmes et pouvoir éventuellement les prédire à partir des évènements passés est aujourd'hui un thème de recherche clé en hydrologie. Diverses études sur les risques de pluies intenses dans la région des Cévennes-Vivarais ont d'ailleurs été réalisées ces dernières années [19, 26, 22, 3]. Elles visaient à cartographier :

- les **temps de retour** de hauteurs d'eau extrêmes pour divers pas de temps (1h, 2h, 6h, 12h, 24h),
- ou inversement, les hauteurs d'eau qui devraient être observées pour un temps de retour donné (c'est ce qu'on appelle un **niveau de retour**).

L'approche considérée dans ces études est la suivante : en chaque station de mesure, on conserve les maxima mensuels et on suppose que leur loi de distribution est une loi de Gumbel [8]. Cette loi dépend de deux paramètres : un paramètre de position et un paramètre d'échelle (appelé par les hydrologues le **gradex**). Ils peuvent être estimés par diverses techniques telles que la méthode des moments, le maximum de vraisemblance ou encore les moindres carrés. Les

temps de retour pour diverses hauteurs de pluies et réciproquement les niveaux de retour pour divers temps de retour se déduisent alors facilement pour chacune des stations. Enfin, l'estimation sur l'ensemble de la région est réalisée par krigeage, technique d'interpolation linéaire issue de la géostatistique [6].
L'ensemble des résultats obtenus par cette approche a été synthétisé sous la forme d'un atlas expérimental [3], fruit d'une collaboration entre le Laboratoire d'étude des Transferts en Hydrologie et Environnement (LTHE) et le laboratoire de la montagne alpine. Une des conclusions principales de cette synthèse est que les risques de forts abats d'eau au pas de 1h ou 2h se situent d'avantage au pied des reliefs alors que pour un pas d'intégration plus élevé (24h), ils s'alignent sur les crêtes.
Toutes les études précédentes ont été réalisées avec une base de données événementielles, c'est à dire que seuls certains épisodes de pluie, jugés importants, ont été enregistrés. De plus, seules les pluies d'automnes ont été considérées. Cette base de données a pour intérêt :

- de fournir une chronique relativement longue (1972 à 1992),
- relativement dense dans l'espace (300 stations réparties sur une fenêtre d'environ $200km \times 200km$),
- d'être au pas de temps horaire.

Ses inconvénients sont les suivants :

- elle est extrêmement hétérogène,
- elle contient finalement peu de mesures puisque seuls certains évènements pluvieux ont été enregistrés,
- elle nous oblige à faire l'hypothèse que toutes les mesures qui n'ont pas été enregistrées sont inférieures à celles retenues pour l'étude de la queue de distribution.

Depuis 1992, environ 150 pluviomètres mesurent en continu la pluie dans la région Cévennes-Vivarais.
Après avoir présenté ces données, nous proposons, dans la première partie de ce rapport, de les utiliser pour refaire une étude similaire aux précédentes (modélisation par station sous hypothèse d'indépendance des mesures + krigeage), l'idée étant de valider les résultats précédents. En particulier, nous nous intéresserons aux deux questions suivantes :

- Est il vraiment justifié de modéliser les valeurs extrêmes par une loi de Gumbel ?
- Les conclusions des études précédentes restent elles les mêmes avec cette nouvelle base de données et avec un nouveau modèle pour les valeurs extrêmes ?

Dans la deuxième partie de ce rapport, nous montrerons que l'hypothèse d'indépendance et de même loi pour les données, sur laquelle repose les calculs précédents, est erronée. Nous nous intéresserons donc au développement d'un modèle prenant en compte la non-stationnarité des mesures. L'approche envisagée repose sur le découpage de l'année en saisons homogènes.

Chapitre 2

Données Cévennes-Vivarais

La base de données de pluie au pas de temps horaire nous a été fournie par le LTHE. Elle contient des mesures horaires entre 1972 à 2000, pour environ 300 stations. C'est une base de données très hétérogène car elle regroupe 3 campagnes de mesures avec des stratégies très différentes :

- Avant 1993, les données sont généralement événementielles et certaines années sont manquantes. Seules les périodes suivantes sont enregistrées : 1972-1980, novembre 1986, octobre 1987 et 1983-1993,
- Entre 1993 et 2000, les mesures ont été enregistrées en continu sur l'année, mais pour seulement 142 stations,
- Après 2000, les mesures ont été enregistrées pour presque 300 stations mais uniquement en automne.

L'ensemble de ces mesures représente 29655 échéances, c'est à dire 29655 heures auxquelles il a plu dans au moins une des stations. L'hétérogénéité de ces mesures, tant d'un point de vue temporel que spatial, rend difficile l'étude des valeurs extrêmes. C'est pourquoi nous avons choisi de travailler uniquement sur une partie de ces données, la plus homogène possible, c'est à dire entre 1993 et 2000. Après avoir retiré les années allant de 1972 à 1993 et de 2001 à 2005, il reste 21881 échéances, ce qui montre à quel point la base de données est vide en dehors de ces années et plus particulièrement avant 1993. Cette nouvelle base reste toutefois hétérogène. D'une part, il reste 7.28% de valeurs manquantes (voir table 2.1) que nous devons supposer inférieures à celles conservées pour notre étude. D'autre part, comme le montre la figure 2.2, qui présente le nombre de mesures positives par mois et par station, ces valeurs manquantes se répartissent différement selon les stations. Globalement, c'est souvent en 1993 qu'il manque des mesures mais il est difficile de généraliser. Si on représente les stations avec un symbole proportionnel au nombre de mesures positives enregistrées entre 1993 et 2000 (voir figure 2.3), on voit que certaines stations devraient être supprimées de l'étude ou aggrégées à la station voisine. Par exemple, à Colombier le jeune, on enregistre 1896 mesures positives, alors que juste à coté, à Colombier le vieux, on en enregistre seulement 477, ce qui est illogique et s'explique par de nombreuses valeurs manquantes. D'autres exemples sont présentés figure 2.3. Comme il est difficile de regarder au cas par cas toutes les incohérences de la base de données, nous avons décidé de supprimer de notre étude toutes les stations avec peu de mesures. Au vu de l'histogramme du nombre de mesures positives par station et par an (figure 2.1) qui montre qu'on a en moyenne 2000 heures de

pluie par station, nous avons décidé d'éliminer toutes les stations avec moins de 1000 mesures positives. Ainsi, sur 142 stations, nous en avons conservé 126, ce qui représente environ une station tous les 11 km (distance moyenne entre une station et sa plus proche voisine). La table 2.2 montre que pour ces stations, il pleut en moyenne 3 mm d'eau par heure avec un écart-type de 3.4 mm et qu'on peut déjà considérer comme relativement extrêmes des pluies de plus de 37 mm par heure puisqu'elles correspondent au percentile 99.9%. Si l'on ramène les données horaires à des données journalières (tableau 2.3), il pleut en moyenne 11.67 mm par jour avec un écart-type de 17.13 mm et on peut considérer comme relativement extrême des pluies supérieures à 160 mm par jour.

Dates manquantes	Données NaN	Valeurs nulles	Valeurs positives
71.17%	7.28%	18.90%	2.65%

TAB. 2.1 – Bilan du nombre de mesures positives, nulles, manquantes ou NaN. Les mesures notées NaN correspondent à des valeurs non mesurées suite par exemple à un pluviomètre défectueux. Les Dates manquantes correspondent à des heures où il n'a plu nulle part. Les valeurs nulles signifie qu'il n'a pas plu dans la station mais qu'il a plus dans au moins une des 141 autres stations.

Moyenne	Ecart-Type	$P_{25\%}$	$P_{50\%}$	$P_{75\%}$	$P_{99\%}$	$P_{99.9\%}$
3.04	3.44	1.2	2	3.5	17.2	36.8

TAB. 2.2 – Statistiques générales sur les données horaires (en mm) : moyenne, écart-type et $x$ièmes percentiles $P_{x\%}$.

Moyenne	Ecart-Type	$P_{25\%}$	$P_{50\%}$	$P_{75\%}$	$P_{99\%}$	$P_{99.9\%}$
11.67	17.13	2	5.5	14.2	83	161.34

TAB. 2.3 – Statistiques générales sur les données journalières (en mm) : moyenne, écart-type et $x$ièmes percentiles $P_{x\%}$.

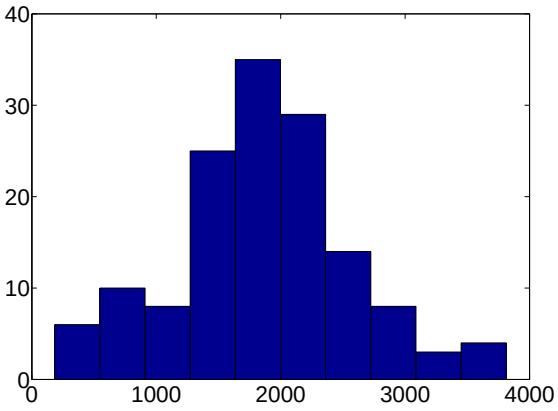

FIG. 2.1 – Histogramme du nombre de mesures positives entre 1993 et 2000 par station

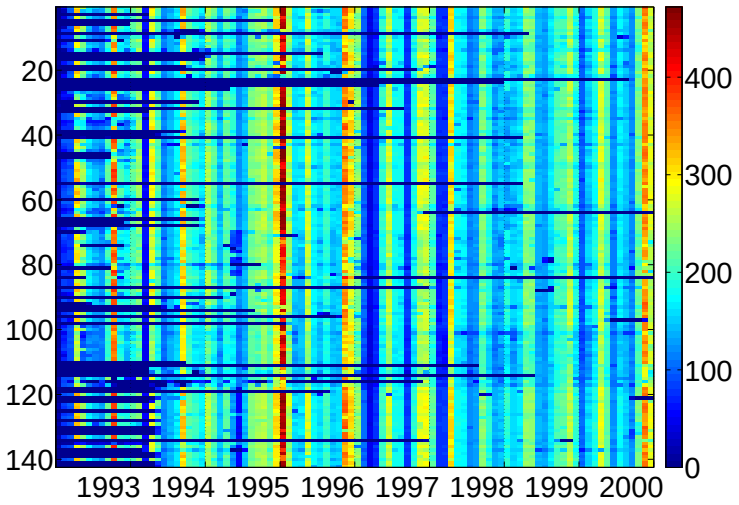

FIG. 2.2 – Nombre de mesures positives par station et par mois

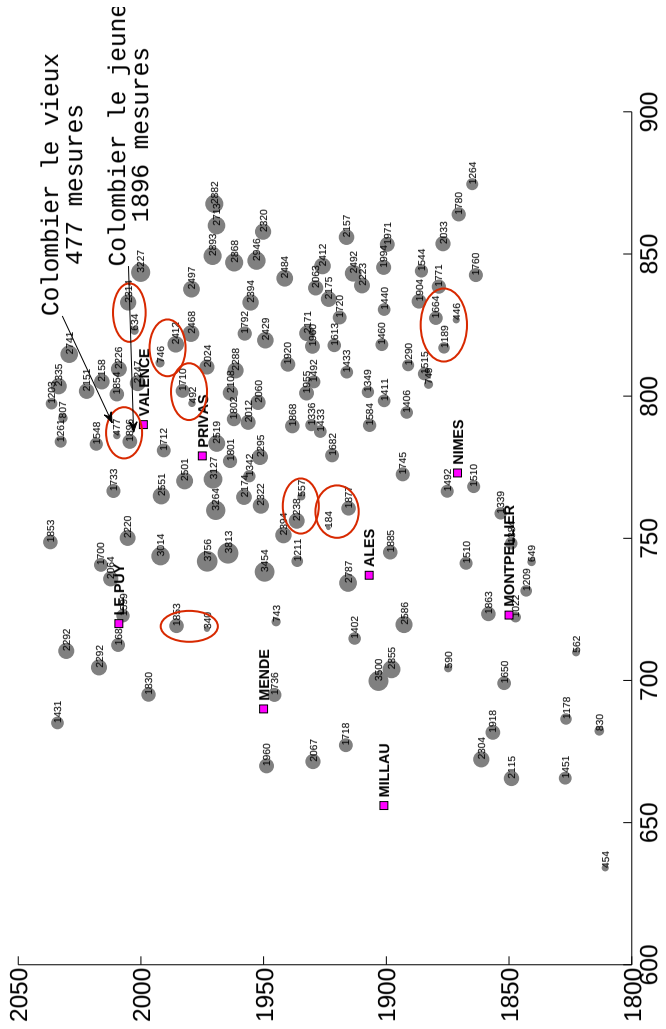

FIG. 2.3 – Nombre de mesures positives par station et par an

Chapitre 3

Analyse des valeurs extrêmes : approche naïve

Dans ce chapitre, nous rappelons brièvement les hypothèses et résultats de la théorie des valeurs extrêmes [8] et de la géostatistique [6]. Ces méthodes sont ensuite appliquées aux mesures de pluie horaires ou journalières pour déterminer les niveaux et temps de retour sur l'ensemble de la région des Cévennes-Vivarais.

3.1 Méthodes

Dans ce paragraphe, nous présentons l'approche POT (Peaks-over-threshold) qui consiste à modéliser la distribution des valeurs dépassant un seuil donné par une loi GPD dont les paramètres seront estimés par maximum de vraisemblance ou par l'estimateur de Hill. Après avoir rappelé comment estimer les temps et niveaux de retour, nous proposerons plusieurs méthodes pour choisir le seuil et le domaine d'attraction de la loi des excès. Enfin, les outils classiques de la géostatistique, variogramme et krigeage, seront introduits.

3.1.1 Etude des excès

Notations-hypothèses

Soit X la variable aléatoire modélisant les hauteurs de pluie horaires (mm) en une station de mesure. On note $\{X_1, \ldots, X_n\}$ l'échantillon de données horaires dont on dispose et $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ l'échantillon ordonné. Dans tout ce chapitre, on supposera les données indépendantes et identiquement distribuées.

Problématique

On s'intéresse à deux problèmes complémentaires :

– Calculer la probabilité d'observer une hauteur de pluie extrême, c'est à dire calculer $p = \mathbb{P}(X > h)$ avec $h > X_{n,n}$. Plus souvent, cette probabilité est exprimée en temps de retour $T = 1/p$. Dans le cas de données horaires, le temps de retour s'exprime en heures. Il doit donc être divisé par le nombre moyen d'heures dans une année $Nh_A = 365.25 * 24$ pour être exprimé en années.

– Calculer la hauteur de pluie h qui est atteinte ou dépassée une seule fois sur T heures avec $T > n$, c'est à dire résoudre $1/T = \mathbb{P}(X > h)$. C'est ce qu'on appelle un niveau de retour.

Pour répondre à ces deux questions, on cherche à modéliser la fonction de survie $\bar{F}(x) = \mathbb{P}(X > x) = 1 - F(x)$ où F est la fonction de répartition de X. On ne cherche pas à la modéliser dans son ensemble, mais uniquement en queue de distribution, c'est à dire quand $x > X_{n,n}$. Deux approches sont possibles :

– La première s'appuie sur un découpage des données en blocs, dont les maxima sont supposés distribués selon une loi de la famille GEV (Generalized Extreme Value, [8], chapitre 3). C'est l'approche typiquement utilisée par les hydrologues, qui supposent de plus que F appartient au domaine de Gumbel.
– La seconde modélise la distribution des valeurs dépassant un seuil donné. On l'appelle la méthode POT (Peaks-over-threshold, [8], chapitre 4). C'est l'approche que nous considérons dans ce rapport, l'approche par maxima n'étant pas envisageable avec seulement 8 années de mesures.

Etude des excès

Dans cette approche, on se fixe un seuil u. On définit alors un excès Y de la variable X au dessus du seuil u par $Y = X - u$ quand $X > u$ (voir figure 3.1). On appelle dépassements les valeurs de X au dessus du seuil u.

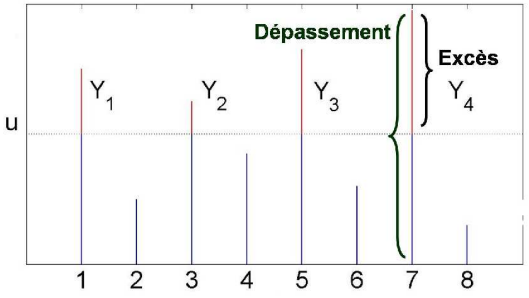

FIG. 3.1 – Définition d'un excès

La fonction de survie d'un excès au dessus de u est donnée pour $y > 0$ par :

$$\begin{aligned}\bar{F}_u(y) &= \mathbb{P}(Y > y) \\ &= \mathbb{P}(X - u > y | X > u) \\ &= \frac{\mathbb{P}(X > u + y, X > u)}{\mathbb{P}(X > u)} \\ &= \frac{\bar{F}(u + y)}{\bar{F}(u)}\end{aligned}$$

Lorsque le seuil est grand, on peut approcher cette quantité par la fonction de survie d'une loi de Pareto Généralisée (GPD) [8] donnée par :

$$\bar{G}_{\gamma,\sigma} = \begin{cases} (1 + \gamma\frac{y}{\sigma})^{-\frac{1}{\gamma}} \text{ si } \gamma \neq 0 \\ \exp(-\frac{y}{\sigma}) \text{ sinon} \end{cases}$$

Son ensemble de définition est $\mathbb{R}^+$ si $\gamma \geq 0$ ou $[0, -\frac{\sigma}{\gamma}[$ si $\gamma < 0$.
La loi GPD dépend de deux paramètres :

- $\sigma > 0$ est le paramètre d'échelle (gradex),
- $\gamma \in \mathbb{R}$ est le paramètre de forme.

On distingue trois types de lois selon la valeur du paramètre de forme γ :

- Le domaine d'attraction de Fréchet quand γ est positif : la fonction de survie décroît comme une puissance de x quand x tend vers l'infini. Ce sont des lois à queue lourde.
- Le domaine d'attraction de Gumbel quand γ est nul : La distribution de X présente une décroissance de type exponentiel dans la queue de distribution. Ce sont des lois à queues légères. Dans la plupart des études sur les pluies ou les débits, les hydrologues supposent que F est dans ce domaine.
- Le domaine d'attraction de Weibull quand γ est négatif : Cela suppose que la distribution de X est bornée, ce qui est peu réaliste dans le cas des pluies. Ce sont les lois à queues finies.

Temps de retour

Connaissant la loi des excès, le temps de retour associé à une hauteur de pluie x se déduit facilement. Comme on a pour $x > u$:

$$\mathbb{P}(X > x | X > u) = \left[1 + \gamma \frac{x-u}{\sigma}\right]^{-\frac{1}{\gamma}}$$

alors

$$\mathbb{P}(X > x) = \xi_u \left[1 + \gamma \frac{x-u}{\sigma}\right]^{-\frac{1}{\gamma}} \text{ si } \gamma \neq 0 \tag{3.1}$$

$$= \xi_u \exp(-\frac{x-u}{\sigma}) \text{ sinon} \tag{3.2}$$

avec $\xi_u = \mathbb{P}(X > u)$.

Niveau de retour

De même, le niveau de retour x_T qui est dépassé en moyenne toutes les T heures est solution de :

$$\xi_u \left[1 + \gamma(\frac{x_T - u}{\sigma})\right]^{-\frac{1}{\gamma}} = \frac{1}{T}.$$

On en déduit :

$$x_T = \begin{cases} u + \frac{\sigma}{\gamma}\left[(T\xi_u)^\gamma - 1\right] \text{ si } \gamma \neq 0 \\ u + \sigma \log(T\xi_u) \text{ sinon} \end{cases} \tag{3.3}$$

Estimation des paramètre de la GPD

Plusieurs méthodes sont possibles pour estimer les paramètres de forme et d'échelle. Nous présentons dans ce rapport deux approches : l'approche par maximum de vraisemblance et l'approche de Hill dans le cas particulier où les données sont dans le domaine de Fréchet.

Dans les applications, nous envisagerons d'abord l'approche par maximum de vraisemblance qui nous permettra de déterminer le domaine d'attraction de la fonction de répartition des excès. On verra que le domaine de Fréchet semble le plus approprié, ce qui nous a conduit à utiliser un deuxième estimateur : l'estimateur de Hill.

Maximum de vraisemblance : Notons $y_1, \ldots, y_k$ les k excès observés au dessus du seuil u.
Pour $\gamma \neq 0$, la log-vraisemblance s'écrit :

$$l(\gamma, \sigma) = \begin{cases} -k \log \sigma - (1 + \frac{1}{\gamma}) \sum_{i=1}^{k} \log(1 + \gamma \frac{y_i}{\sigma}) \text{ si } (1 + \sigma^{-1} \gamma y_i) > 0 \text{ pour } i = 1, \ldots, k \\ -\infty \text{ sinon} \end{cases}$$

Pour $\gamma = 0$, elle s'écrit :

$$l(\gamma, \sigma) = -k \log \sigma - \sigma^{-1} \sum_{i=1}^{k} y_i.$$

Les paramètres de la loi GPD (γ et σ) sont alors estimés en maximisant la log-vraisemblance. Comme il n'existe pas de solution explicite, le recours à des méthodes numériques est indispensable. Il est possible de fournir des intervalles de confiance asymptotiques pour les paramètres de la loi GPD. Pour plus de détails, voir [8], chapitre 2, pages 32-33.

Estimateur de Hill : Lorsqu'on se restreint au domaine de Fréchet, on a la caractérisation :

$$\bar{F}(x) = x^{-\frac{1}{\gamma}} l(x)$$

avec $\gamma > 0$ et l une fonction à variations lentes, c'est à dire que pour $t > 1$

$$\lim_{u \to \infty} \frac{l(tu)}{l(u)} = 1$$

ce qui conduit à :

$$\lim_{u \to \infty} \frac{\bar{F}(tu)}{\bar{F}(u)} = t^{-\frac{1}{\gamma}}$$

ou encore, quand u est grand :

$$\bar{F}(tu) \simeq t^{-\frac{1}{\gamma}} \bar{F}(u)$$

En posant $x = tu$, on a donc quand u est grand :

$$\bar{F}(x) \simeq \bar{F}(u)(\frac{x}{u})^{-\frac{1}{\gamma}} \simeq \xi_u (\frac{x}{u})^{-\frac{1}{\gamma}} \tag{3.4}$$

Quand p est proche de zéro, on a aussi :

$$\bar{F}^{-1}(p) \simeq \bar{F}^{-1}(\xi_u)(\frac{p}{\xi_u})^{-\gamma}. \tag{3.5}$$

On en déduit

$$\log \bar{F}^{-1}(p) - \log \bar{F}^{-1}(\xi_u) \simeq \gamma \log(\frac{\xi_u}{p})$$

En posant $\xi_u = k/n$, avec $k \in 1, \ldots, n-1$ et en choisissant plusieurs valeurs de $p = i/n$ avec $i = 1, \ldots, k$, on obtient :

$$\log \bar{F}^{-1}(\frac{i}{n}) - \log \bar{F}^{-1}(\frac{k}{n}) \simeq \gamma \log(\frac{k}{i})$$

ou encore en estimant les fonctions de survie par leurs équivalents empiriques :

$$\log X_{n-i,n} - \log X_{n-k,n} \simeq \gamma \log(\frac{k}{i})$$

Cette approximation peut être vérifiée graphiquement en traçant $\log X_{n-i,n} - \log X_{n-k,n}$ en fonction de $\log(k/i)$. Ce graphique est appelé **diagramme de Hill**. Il permet de vérifier qu'on est bien dans le domaine d'attraction de Fréchet auquel cas le diagramme de Hill est une droite dont la pente correspond au paramètre de forme γ. Si on est bien dans le domaine de Fréchet, on peut alors estimer γ par

$$\hat{\gamma}(k) = \frac{1}{k} \sum_{i=1}^{k-1} (\log X_{n-i,n} - \log X_{n-k,n})$$

Cet estimateur est appelé estimateur de Hill.
On peut associer à l'estimateur de Hill un intervalle de confiance asymptotique $I_N(\alpha)$ de niveau α.

$$I_N(\alpha) = [\hat{\gamma} - z_{\alpha/2}\hat{\gamma}\frac{1}{\sqrt{k}}, \hat{\gamma} + z_{\alpha/2}\hat{\gamma}\frac{1}{\sqrt{k}}]$$

où $z_{\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite. Pour plus de détails, voir [14], page 74. Remarque : Dans le domaine de Fréchet, le paramètre d'échelle s'exprime en fonction du paramètre de forme par la relation $\sigma = u\gamma$. En remplaçant σ par cette expression dans les équations (3.1) et (3.3), on retrouve bien les équations générales des temps et niveaux de retour (3.4) et (3.5).

Choix du seuil

Plutôt que de se fixer un seuil u de pluie pour chacune des stations, nous avons choisi de fixer le nombre de valeurs fortes i que nous souhaitions conserver. Si $X_1, \ldots, X_n$ est l'échantillon de taille n et que l'on souhaite garder les k plus grandes valeurs de l'échantillon, alors le seuil est estimé par $u = X_{n-k,n}$ et la probabilité de dépasser ce seuil par $\xi_u = \mathbb{P}(X > u) \approx k/n$. Ceci est valable lorsque les données sont toutes distinctes, ce qui n'est pas le cas des données de pluie pour lesquelles on observe de nombreuses fois une même valeur. Dans ce cas, nous proposons d'estimer le seuil par :

$$u = \begin{cases} X_{n-k,n} \text{ si } X_{n-k,n} \neq X_{n-k+1,n} \\ X_{n-k-i,n} \text{ si } \forall j = 0, \ldots, i-1 \ , \ X_{n-k-j,n} = X_{n-k+1,n} \text{ et } X_{n-k-i,n} \neq X_{n-k+1,n} \end{cases}$$

et la probabilité de dépasser ce seuil par :

$$\mathbb{P}(X > u) = \frac{k+i}{n}$$

Le nombre de valeurs dépassant le seuil n'est donc plus k mais $k + i$. Le choix du seuil ou du nombre d'excès est crucial car il peut radicalement changer les

estimations et leurs intervalles de confiance. Si le nombre d'excès $k + i$ que l'on conserve est trop petit (ou de manière équivalente, le seuil u trop grand), alors on a de grandes chances que les excès suivent bien une loi GPD, mais les estimations seront fortement instables car le nombre de mesures ne sera pas suffisant. A l'inverse, si $k + i$ est trop grand (ou de manière équivalent u petit), alors l'hypothèse d'une loi GPD sur les excès ne sera plus vérifiée. Les estimations seront plus stables mais biaisées. Il faut donc trouver un seuil tel qu'on ait suffisamment de mesures tout en restant dans l'hypothèse d'une loi GPD pour les excès. Pour choisir le nombre d'excès, nous proposons trois méthodes :

- Tracer, en fonction du nombre d'excès (exprimé en pourcentage), les estimations des paramètres de forme et d'échelle et regarder si on distingue une zone pour laquelle les estimations sont stables.
- Comparer les estimations par maximum de vraisemblance et par Hill des paramètres de forme et d'échelle et regarder pour quel seuil maximal ces estimations sont cohérentes.
- Réaliser un test d'adéquation à la loi GPD et tracer la p-valeur en fonction du nombre d'excès. On choisit le nombre d'excès maximal tel que la p-valeur soit supérieure à 0.05. On peut utiliser par exemple le test du Chi2 ou le test d'Anderson Darling. ([20], chapitre 7)
  A noter : Lorsque l'on est dans le domaine de Fréchet, alors les variables

$$Z_i = \log \frac{Y_i}{u}$$

  où Y_i sont les excès, suivent approximativement une loi exponentielle de paramètre γ. On peut donc, au lieu de réaliser un test d'adéquation des excès à la loi GPD, réaliser un test d'adéquation des variables Z_i à la loi exponentielle.

Choix du domaine d'attraction

L'estimation du paramètre de forme γ par maximum de vraisemblance permet dans un premier temps de voir si le domaine d'attraction semble être le même pour toutes les stations ou si au contraire il change d'une région à l'autre. Il est intéressant par ailleurs de regarder si il y a bien une cohérence spatiale dans les estimations. A priori, il ne semble pas réaliste d'avoir des estimations négatives de γ car il n'y a aucune raison pour que les hauteurs de pluie soient bornées. Le domaine de Gumbel et de Fréchet semblent donc plus adaptés à l'étude des pluies. Une deuxième approche intéressante pour déterminer le domaine d'attraction le plus réaliste est d'étudier les estimations de γ par l'estimateur de Hill. Si ces estimations sont très proches de zéro, c'est qu'on est probablement dans le domaine de Gumbel, si au contraire elles s'en éloignent, on est plutôt dans le domaine de Fréchet. Enfin, des outils graphiques tels que le diagramme de Hill (décrit paragraphe 3.1.1) ou le return level plot ([8], page 49) peuvent aussi être utilisés pour choisir le domaine d'attraction.

Return level plot : On considère les maxima $m_1, \ldots, m_T$ de T blocs (chaque bloc est par exemple une année ou un mois). On trace ces maxima en fonction de

$$-\log(-\log(\hat{F}(m_i))), i \in \{1, \ldots, T\}.$$

où $\hat{F}$ est la fonction de répartition empirique de X définie par :

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x) \tag{3.6}$$

Alors :
- Si on est dans le domaine d'attraction de Gumbel, la courbe est linéaire,
- Si la courbe est convexe avec une asymptote horizontale, on est dans le domaine d'attraction de Weibull,
- Et enfin, si la courbe est concave et non bornée, on est dans le domaine d'attraction de Fréchet.

3.1.2 Géostatistique

En géostatistique, on considère que la variable régionalisée $z(x)$ (ici la hauteur de pluie) en tout point x du champs $\mathcal{D}$ étudié est la réalisation d'une fonction aléatoire $Z(x)$. Le nombre d'observations disponibles étant limité, il est illusoire de vouloir inférer la loi spatiale entière de Z et on se restreint à l'étude des deux premiers moments et plus particulièrement à l'étude du variogramme qui caractérise la corrélation spatiale entre sites. Ce variogramme est ensuite utilisé dans l'étape de krigeage, technique d'interpolation linéaire permettant d'estimer les hauteurs de pluies entre les sites de mesures [6].

Le variogramme

N'ayant en général qu'une réalisation de la fonction aléatoire Z en chaque site de mesure, l'inférence est impossible sauf si l'on se restreint à l'étude de fonctions aléatoires stationnaires (ou intrinsèques), c'est à dire que la loi spatiale de la fonction aléatoire (ou de ses accroissements) est invariante par translation. En d'autres termes, l'espérance et la variance de la différence $Z(x+h) - Z(x)$ entre deux sites ne dépend que de la distance h qui les sépare.
On suppose donc que $Z(x)$ est une fonction aléatoire intrinsèque sans dérive :

$$\forall x, x+h \in \mathcal{D}, \begin{cases} E[Z(x+h) - Z(x)] = 0 \\ var[Z(x+h) - Z(x)] = 2V(h) \end{cases}$$

où $V(h)$ est appelé variogramme et $\mathcal{D}$ est le champs d'étude (domaine borné de $\mathbb{R}^d$).
Le variogramme mesure la variabilité des mesures entre deux points séparés par une distance h. Souvent, $V(h)$ croît à partir de $h = 0$, puis atteint, à partir d'une distance a (la portée), une valeur limite (le palier, voir figure 3.2). Cela signifie que lorsque deux points sont séparés par une distance supérieure à la portée, les variables aléatoires associées à ces points ne sont plus corrélées.
Soient $z(x_i), x_i \in \mathcal{D}, i \in \{1, \ldots, n\}$ les données expérimentales, on définit alors un estimateur du variogramme de la manière suivante :

$$\hat{V}(h) = \frac{1}{2|N(h)|}\sum_{N(h)} [z(x_\alpha) - z(x_\beta)]^2$$

où $N(h) = \{(\alpha, \beta) \text{ tel que } x_\alpha - x_\beta = h\}$ et $|N(h)|$ est le nombre de paires distinctes de l'ensemble $N(h)$.

L'estimateur ainsi défini est appelé variogramme expérimental. On lui ajuste ensuite un modèle (variogramme sphérique, exponentiel ...) [6] qui permet de calculer la valeur du variogramme pour n'importe quelle distance, ce qui est indispensable pour résoudre les équations de krigeage. Dans cette étude, nous utiliserons principalement les modèles pépitiques et sphériques décrits ci-dessous :

– le modèle pépitique de palier C (appelé aussi effet de pépite) traduit une absence de structure spatiale et quantifie les erreurs de mesure,

$$V(h) = \begin{cases} 0 \text{ pour } h = 0 \\ C \text{ pour } h > 0 \end{cases}$$

– le modèle sphérique de portée a et de palier C traduit un comportement linéaire à l'origine

$$V(h) = \begin{cases} C \text{ pour } h \geq a \\ C(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\frac{h^3}{a^3}) \text{ pour } 0 \leq h \leq a \end{cases}$$

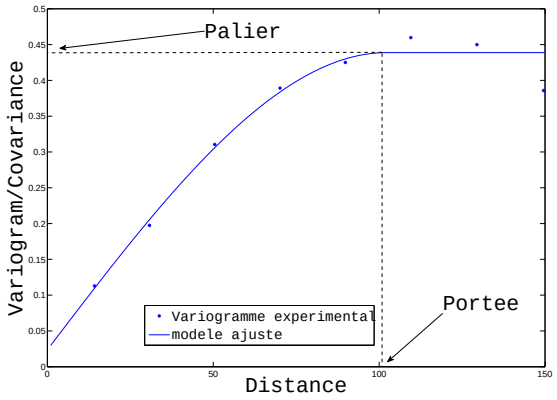

FIG. 3.2 – Exemple de variogramme expérimental modélisé par la somme d'un effet de pépite et un variogramme sphérique

Le krigeage

Le krigeage permet d'estimer la hauteur de pluie en un point à partir des valeurs observées sur les sites voisins. Il prend en compte la configuration géométrique de points et leur structure spatiale via l'utilisation du variogramme. Le krigeage est défini par les 4 contraintes suivantes :

– Contrainte de linéarité : on estime la valeur en un point x_0 par une combinaison linéaire des valeurs aux sites voisins $x_1, \ldots, x_n$.

$$\hat{Z}(x_0) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(x_\alpha)$$

– Contrainte d'autorisation : toute combinaison linéaire des $Z(x_i), i \in \{1, \ldots, n\}$ possède une espérance et une variance finies. Dans le cadre stationnaire, cette condition est toujours vérifiée. Dans le cadre intrinsèque, on doit imposer que la somme des poids soit 1.

$$\sum_{\alpha=1}^{n} \lambda_\alpha = 1$$

– Contrainte de non biais :

$$E(\hat{Z}(x_0) - Z(x_0)) = 0$$

elle conduit à imposer que la somme des poids soit 1.

– Contrainte d'optimalité : on cherche les poids qui minimisent la variance d'estimation $Var(\hat{Z}(x_0) - Z(x_0))$

On montre facilement que ce problème de minimisation sous contrainte revient à résoudre le système suivant :

$$\begin{pmatrix} \gamma(x_1 - x_1) & \ldots & \gamma(x_1 - x_n) & 1 \\ \vdots & & & \vdots \\ \gamma(x_n - x_1) & \ldots & \gamma(x_n - x_n) & 1 \\ 1 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix}$$

où μ est le paramètre de Lagrange.
On en déduit alors facilement l'estimation de la variable régionalisée en x_0 :

$$\hat{Z}(x_0) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(x_\alpha)$$

et la variance de krigeage associée

$$Var(\hat{Z}(x_0) - Z(x_0)) = \sum_{\alpha=1}^{n} \lambda_\alpha \gamma(x_\alpha - x_0) - \mu.$$

3.2 Première analyse des valeurs fortes

Dans cette section, nous présentons l'application des méthodes précédentes aux données de la région Cévennes-Vivarais. Dans un premier temps nous regardons quel domaine d'attraction semble le plus pertinent pour les valeurs fortes. Nous présentons ensuite les cartes de niveaux de retour et de période de retour. Deux pas de temps sont étudiés pour les hauteurs de pluies : le cumul de pluie horaire ou journalier.

3.2.1 Domaine d'attraction ?

Une première approche pour déterminer quel domaine d'attraction semble le plus pertinent pour les hauteurs de pluie extrêmes est d'estimer les paramètres de la loi GPD par maximum de vraisemblance pour toutes les stations. On réalise ensuite un krigeage pour estimer le paramètre de forme sur la région entière. Cela permet de voir si les estimations sont cohérentes spatialement et

si le comportement des valeurs fortes diffère d'une région à l'autre. Rappelons que si le paramètre de forme est positif, on est dans le domaine de Fréchet, si il est proche de zéro, on est dans le domaine de Gumbel et si il est négatif on est dans le domaine de Weibull, que nous jugeons peu réaliste pour l'étude des hauteurs de pluie.
Une deuxième approche pour déterminer le domaine d'attraction consiste à tracer le diagramme de Hill et le return level plot (section 3.1.1). Seules les stations de Barnas, Mazan-L'Abbaye et Valleraugue seront étudiées. Ce sont les stations de mesure les mieux informées du réseau.
Enfin, comme l'hypothèse du domaine d'attraction de Weibull semble peu probable, on peut supposer que l'on est dans le domaine de Fréchet et estimer le paramètre de forme par l'estimateur de Hill. Si les estimations sont éloignées de zéro, l'hypothèse du domaine de Fréchet sera vérifiée, si en revanche elles sont proches de zéro, le domaine de Gumbel sera peut être plus approprié.

Pas de temps horaire

Dans cette sous-section, les résultats sont présentés pour des cumuls de pluie horaires. L'estimation du paramètre de forme par maximum de vraisemblance (voir figure 3.4) montre que ce dernier est positif (> 0.14) sur l'ensemble de la région : Le domaine de Fréchet semble donc le plus adapté. Pour les stations les mieux informées, le diagramme de Hill confirme l'hypothèse du domaine de Fréchet car on obtient bien une droite pour les 3 stations (figure 3.3, colonne droite). Les return level plot pour ces trois mêmes stations indiquent plutôt le domaine de Weibull à Barnas et Valleraugue et éventuellement le domaine de Gumbel à Mazan-L'Abbaye (figure 3.3, colonne gauche). Ces derniers résultats sont surprenants car ils ne coïncident pas avec les estimations du paramètre de forme par maximum de vraisemblance. Notons toutefois que les return level plot sont tracés pour seulement 7 maxima annuels, ce qui n'est pas très fiable. Enfin, si l'on suppose qu'on est effectivement dans le domaine de Fréchet, l'estimation du paramètre de forme par l'estimateur de Hill (figure 3.6) montre que ce dernier est toujours supérieur à 0.32 ce qui confirme que l'ajustement de la queue de distribution par une loi de Gumbel n'est pas très adapté.

Globalement, les estimations sont cohérentes spatialement pour les deux approches : maximum de vraisemblance et Hill (figures 3.4 et 3.6). La cartographie du paramètre de forme obtenue par maximum de vraisemblance est très différente de celle obtenue par Hill. Dans le cas de l'estimateur de Hill, le paramètre de forme semble très fortement lié au relief (voir figure B.1 en annexe B), avec des valeurs fortes en plaine entre les Alpes et le Massif Central, ce que confirme la figure 3.8 qui présente le paramètre de forme estimé par station en fonction de l'altitude. Dans le cas de l'estimation par maximum de vraisemblance, il n'y a pas particulièrement de logique dans les estimations.
Nous présentons figures 3.5 et 3.7 les cartes de variance de krigeage pour les deux méthodes. On ne peut les utiliser telles quelles pour donner des incertitudes sur l'estimation du paramètre de forme mais elles permettent de distinguer les zones pour lesquelles les estimations sont peu fiables (celles avec une grande variance de krigeage). Ce sont bien évidemment les zones qui contiennent peu ou pas de données. Par la suite, nous ne présenterons plus les cartes de variances de krigeage, qui sont pour la plupart identiques, mais nous retiendrons que les estimations dans les zones sans mesure sont peu fiables.

Pour les deux approches, les estimations du paramètre d'échelle sont similaires avec des valeurs toutefois légèrement plus faibles pour l'estimateur de Hill. Dans les deux cas, le paramètre d'échelle semble toujours lié au relief présentant des valeurs fortes sur les plaines et moins élevées en montagne. Toutefois, sa décroissance en fonction de l'altitude est nettement moins évidente que pour le paramètre de forme (voir figure 3.11).
L'ensemble de ces résultats a été obtenu pour un pourcentage d'excès fixé à 7%, ce qui correspond à un seuil moyen de 7mm et à en moyenne 148 mesures par station. Ce pourcentage a été choisi selon les recommandations de la section 3.1.1. Les résultats sont présentés en annexe C. Ils montrent en fait qu'un seuil pertinent pour l'étude serait de 4%, ce qui correspont à un seuil moyen de 9 mm et environ une moyenne de 84 mesures par station. Le graphique 3.12 présente les estimations du paramètre de forme, pour la station de Barnas, par maximum de vraisemblance et par l'estimateur de Hill en fonction du pourcentage d'excès retenus. Les deux estimateurs indiquent que le paramètre de forme est toujours positif, il y a donc cohérence entre les deux méthodes d'estimation pour le choix du domaine d'attraction. En revanche, pour plus de 5% d'excès les intervalles de confiance ne se recoupent plus, ce qui n'est pas logique et confirme le choix du pourcentage d'excès à 4%. Cependant, avec 4%, le nombre de mesures semble insuffisant pour une étude par maximum de vraisemblance. Avec moins de 7% d'excès, les variogrammes pour les paramètres de la loi GPD ne montrent pas de cohérence spatiale, ce qui n'est pas très réaliste. Nous avons donc choisi de fixer le nombre d'excès à 7% pour les deux approches : maximum de vraisemblance et Hill.

Le choix du seuil est extrêmement important car les estimations, les intervalles de confiance, la validité du modèle mais aussi l'analyse finale des résultats en dépendent. En annexe E, les estimations des paramètres de forme et d'échelle, par maximum de vraisemblance ou par l'estimateur de Hill, sont présentées en fonction du pourcentage d'excès retenus. On peut voir que les cartes d'estimation du paramètre de forme sont très différentes en fonction du pourcentage retenu, en particulier pour l'approche par maximum de vraisemblance. Par maximum de vraisemblance, lorsque le pourcentage d'excès est faible, il apparaît une zone de valeurs fortes à l'ouest et de valeurs faibles à l'est. En augmentant fortement le nombre d'excès (à partir de 20% d'excès), on retrouve la zone de valeurs fortes en plaine entre les montagnes, dans la zone englobant Alès, Nîmes et Montpellier. Les cartes obtenues par l'estimateur de Hill diffèrent elles aussi en fonction du pourcentage d'excès mais restent toutefois plus cohérentes entre elles avec une zone de valeurs fortes toujours dans la même zone, en plaine. A l'inverse, par maximum de vraisemblance, les estimations restent toujours dans la même gamme de valeurs (entre 0.1 et 0.28) alors qu'avec l'estimateur de Hill, elles croissent lorsque le seuil croît.

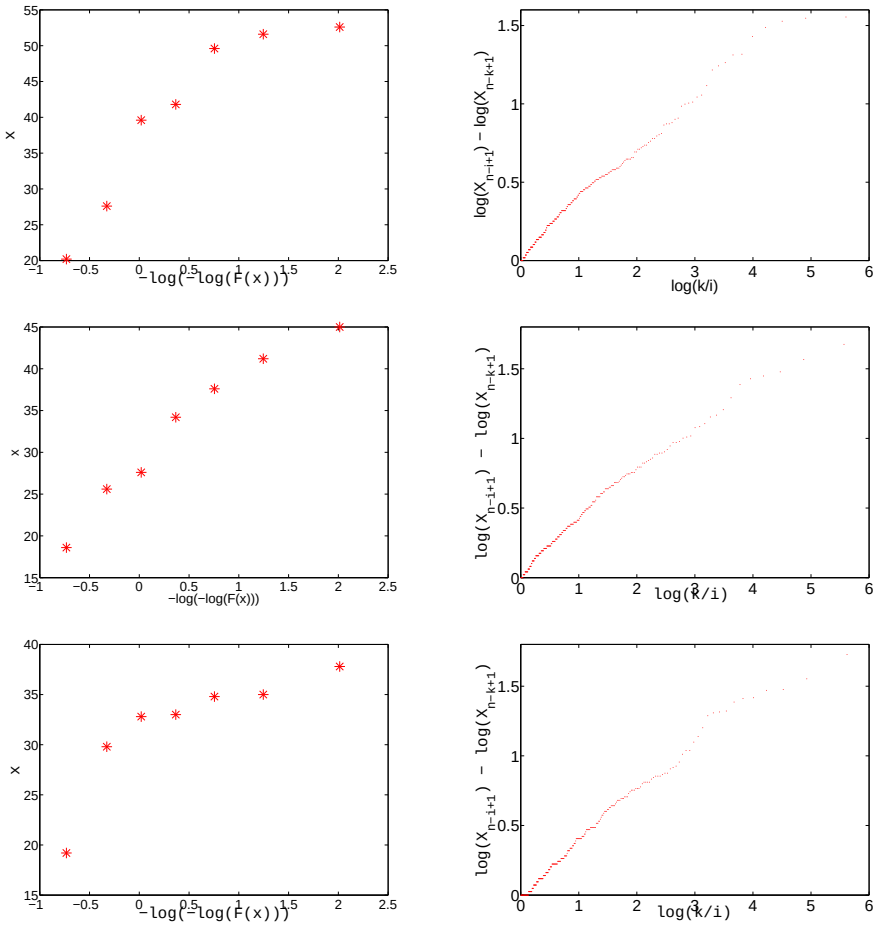

FIG. 3.3 – A gauche : Return level plot. A droite : Diagramme de Hill. Stations de mesures étudiées : Barnas (en haut), Marzan-L'Abbaye (au milieu) et Valleraugue (en bas). Pas de temps horaire.

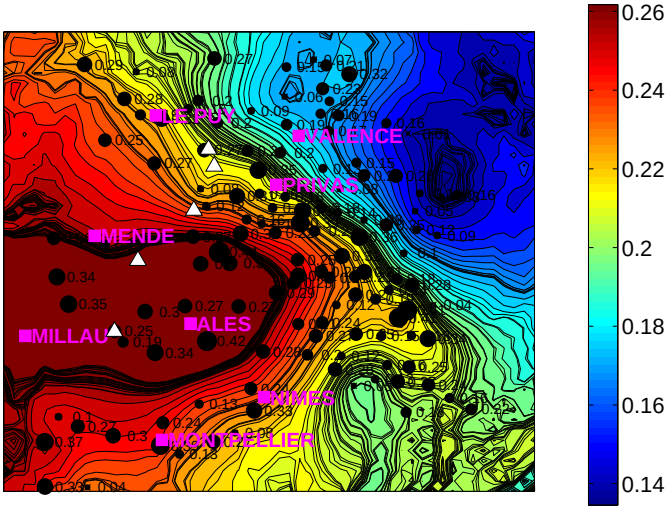

FIG. 3.4 – Estimation par station du paramètre de forme γ par maximum de vraisemblance puis interpolation par krigeage. Pas de temps horaire.

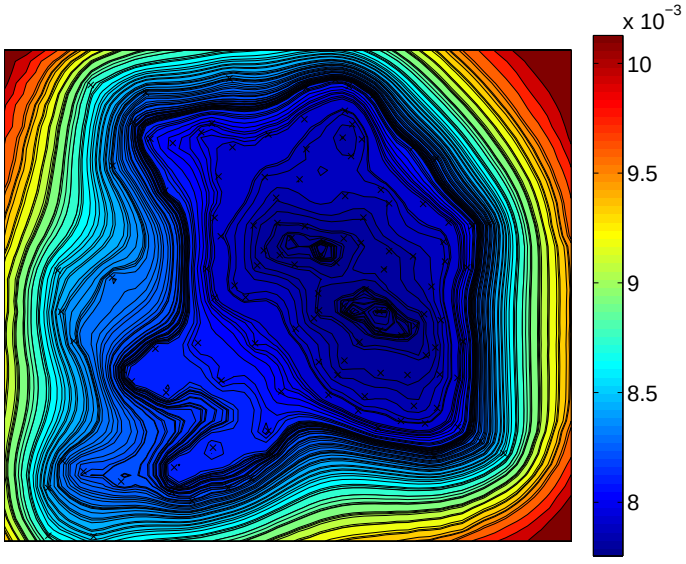

FIG. 3.5 – Variance de krigeage pour l'estimation du paramètre de forme γ par maximum de vraisemblance. Pas de temps horaire.

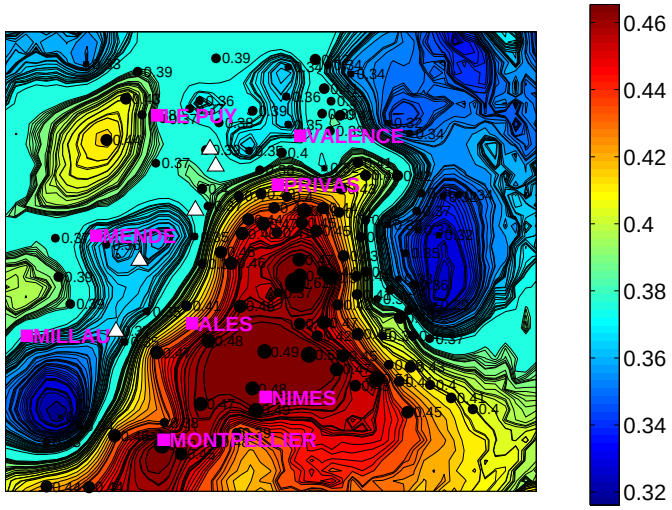

FIG. 3.6 – Estimation par station du paramètre de forme γ par l'estimateur de Hill puis interpolation par krigeage. Pas de temps horaire

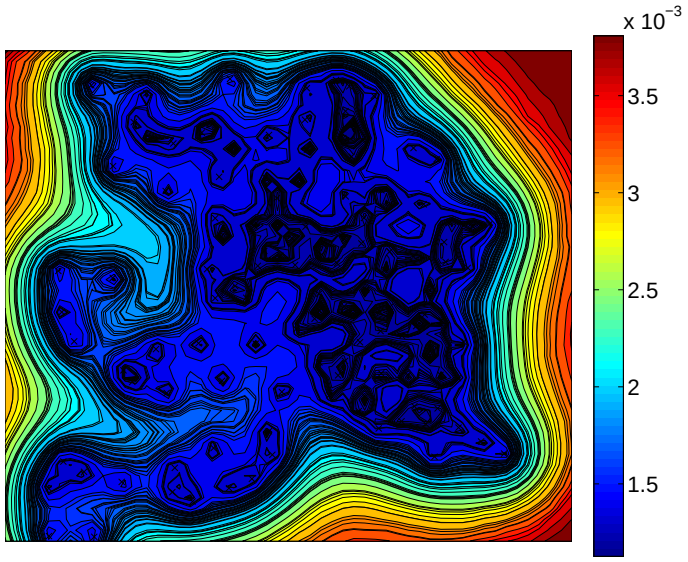

FIG. 3.7 – Variance de krigeage pour l'estimation du paramètre de forme γ par l'estimateur de Hill. Pas de temps horaire

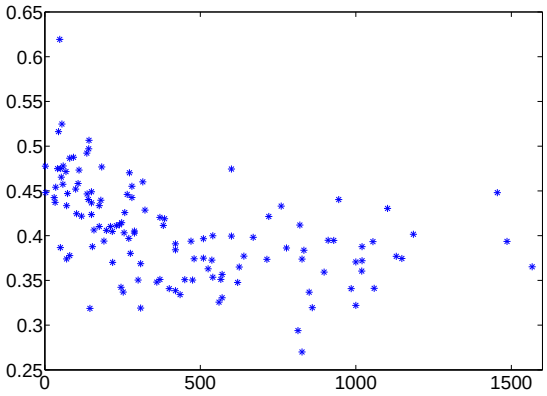

FIG. 3.8 – Paramètre de forme estimé par Hill en fonction de l'altitude. Pas de temps horaire.

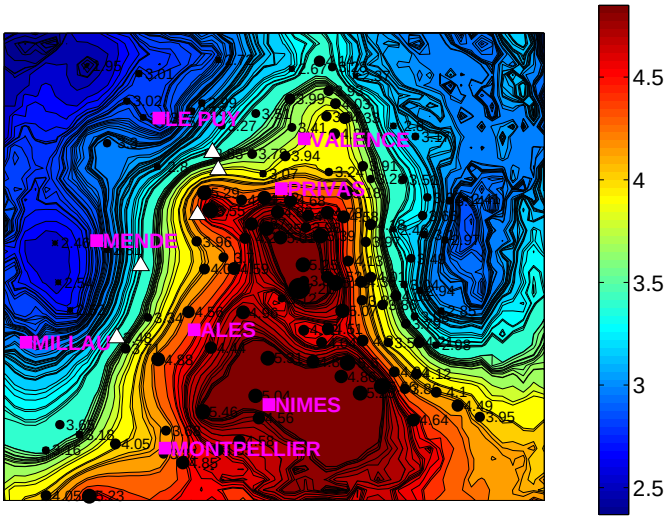

FIG. 3.9 – Estimation par station du paramètre d'échelle σ par maximum de vraisemblance puis interpolation par krigeage. Pas de temps horaire.

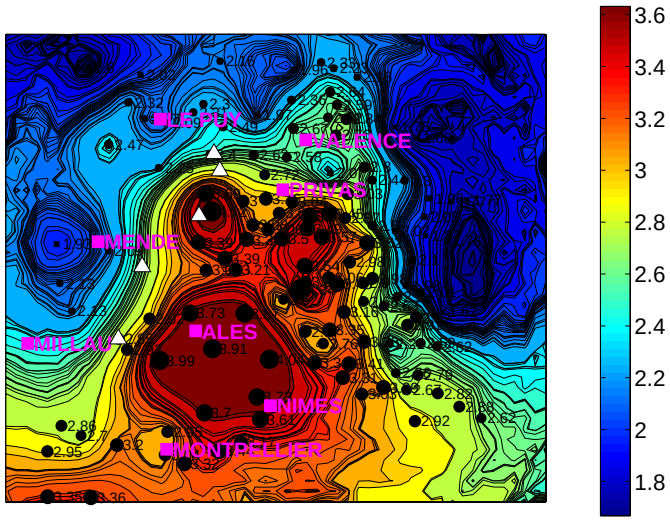

FIG. 3.10 – Estimation par station du paramètre d'échelle σ dans le cas particulier du domaine d'attraction de Fréchet pour lequel $\sigma = u \times \gamma$ où u correspond au seuil et γ au paramètre de forme. Le paramètre de forme γ est estimé par l'estimateur de Hill. Pas de temps horaire.

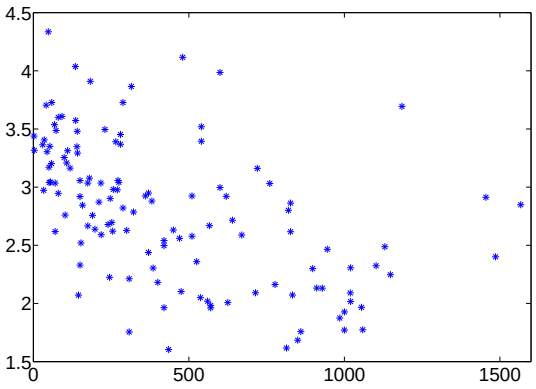

FIG. 3.11 – Paramètre d'échelle estimé par Hill en fonction de l'altitude. Pas de temps horaire.

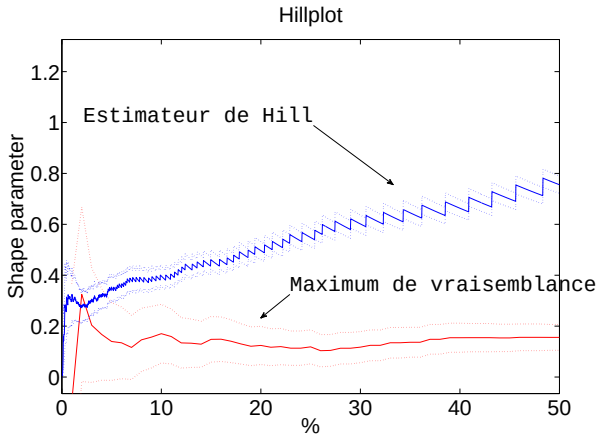

FIG. 3.12 – Station de mesure à Barnas : Paramètre de forme estimé par maximum de vraisemblance et par l'estimateur de Hill en fonction du pourcentage d'excès. Données horaires.

Pas de temps journalier

Lorsque le cumul de pluie est étudié sur 24 heures, c'est à dire au pas de temps journalier, les conclusions diffèrent complètement. Notons tout d'abord qu'en ramenant les données horaires à des données journalières, le nombre de mesures diminue fortement. Or il est montré en annexe D qu'un choix pertinent pour le pourcentage d'excès à retenir est de 10%, ce qui correspond à un seuil de 30 mm de pluie par jour et 52 mesures en moyenne par station, ce qui est très peu. Une estimation des paramètres de la loi GPD par maximum de vraisemblance conduit alors à de nombreuses valeurs négatives pour le paramètre de forme, ce qui n'est pas logique car cela signifie que la hauteur de pluie journalière est bornée pour de nombreuses stations. Les périodes de retour estimées pour un cumul de pluie fixé à 200mm par jour sont alors énormes voir infinies, ce qui signifie que dans certaines régions il ne pleuvra jamais plus de 200mm. Est ce vraiment réaliste ? Les return level plot et diagrammes de Hill pour les trois stations les mieux informées semblent d'ailleurs indiquer que les domaines de Fréchet ou Gumbel sont les plus pertinents (figure 3.19).
Par ailleurs, si on compare les estimations du paramètre de forme par maximum de vraisemblance et par l'estimateur de Hill en fonction du pourcentage d'excès (exemple à Barnas figure 3.14), on voit que les intervalles de confiance de chaque estimateur ne se recoupent quasi jamais sauf pour moins de 10% d'excès. Or avec moins de 10% d'excès, les incertitudes par maximum de vraisemblance sont tellement fortes qu'on ne peut déterminer le domaine d'attraction, le paramètre de forme variant entre des valeurs négatives et des valeurs positives. Avec 10% d'excès, une approche par maximum de vraisemblance ne semble pas pertinente et nous avons donc choisi d'augmenter le pourcentage d'excès à 30%, ayant toutefois bien conscience qu'avec 30%, l'hypothèse d'une loi GPD pour les excès semble compromise. Nous avons cependant conservé 10% d'excès pour l'approche Hill. Notons que 30% d'excès correspondent à un seuil de 12 mm de pluie par jour et environ 159 mesures par station. Les cartographies du paramètre de forme obtenues par maximum de vraisemblance (figure 3.15) et par l'estimateur de Hill (figure 3.16) diffèrent. Par maximum de vraisemblance, on distingue les valeurs fortes dans l'ouest de la région et des valeurs proches de zéro à l'est. Les estimations ne semblent pas liées au relief et sont comprises entre 0 et 0.15. Par l'approche Hill, les estimations sont nettement plus élevées et varient entre 0.34 et 0.5, ce qui indique comme pour les données horaires qu'on est dans le domaine de Fréchet. Ces estimations semblent liées aux relief avec des valeurs fortes sur les Cévennes et des valeurs plus faibles en plaine. Cependant, l'altitude n'explique pas à elle seule le paramètre de forme puisque à l'est, près des Alpes, ce dernier ne semble pas plus élevé qu'ailleurs. Par conséquent, tracer le paramètre de forme en fonction de l'altitude ne laisse apparaître aucune relation entre les deux (figure 3.13). Pour visualiser les altitudes des stations, le lecteur pourra se reporter à l'annexe B, figures B.1 et B.2.
Les cartographies du paramètre d'échelle sont par contre très similaires pour les deux approches. Dans les deux cas, le paramètre d'échelle varie entre 6 et 22. On distingue trois zones :

- une zone de valeurs fortes sur la ligne de crêtes des Cévennes
- une zone de valeurs faibles dans les Alpes
- une zone de valeurs intermédiaires en plaine

Synthèse des résultats

Les principales conclusions sur l'analyse station par station des paramètres de la loi GPD sont les suivantes :

- Quel que soit le pas de temps considéré, horaire ou journalier, la loi des excès semble appartenir au domaine de Fréchet.
- Quel que soit le pas de temps considéré, horaire ou journalier, la cartographie du paramètre d'échelle est similaire pour les deux approches, maximum de vraisemblance et Hill. Au pas de temps horaire, le paramètre d'échelle semble plus élevé en plaine. Au pas de temps journalier, on distingue une zone de valeurs fortes sur les Cévennes, une zone de valeurs intermédiaires en plaine et un zone de valeurs faibles sur les Alpes.
- En revanche, la cartographie du paramètre de forme diffère selon la méthode d'estimation choisie :
  - Au pas de temps horaire et dans le cas de l'estimateur de Hill, le paramètre de forme semble lié au relief, avec des valeurs fortes en basse altitude. Dans le cas du maximum de vraisemblance, la carte est très différente et il ne semble pas y avoir de logique dans les estimations. Le choix du seuil semble très important, en particulier pour les estimations par maximum de vraisemblance. Avec peu de mesures, les incertitudes sur les estimations par maximum de vraisemblance sont très fortes et l'approche par estimateur de Hill nous semble donc plus pertinente, d'autant plus que l'hypothèse du domaine de Fréchet pour la loi des excès est confirmée pas les deux approches.
  - Au pas de temps journalier et dans le cas de l'estimateur de Hill, on distingue pour le paramètre de forme une zone de valeurs fortes sur les Cévennes. Il ne semble donc pas uniquement lié à l'altitude puisque les valeurs ne sont pas particulièrement élevées dans les Alpes. Dans le cas du maximum de vraisemblance, la carte est encore une fois très différente et il ne semble pas y avoir de logique dans les estimations.

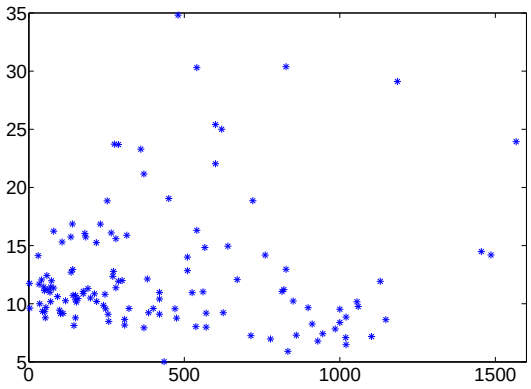

FIG. 3.13 – Paramètre de forme estimé par Hill en fonction de l'altitude. Données journalières.

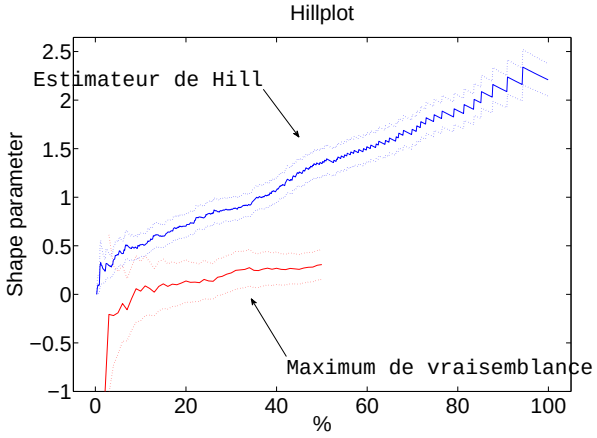

FIG. 3.14 – Station de mesure à Barnas : Paramètre de forme estimé par maximum de vraisemblance et par l'estimateur de Hill en fonction du pourcentage d'excès. Données journalières.

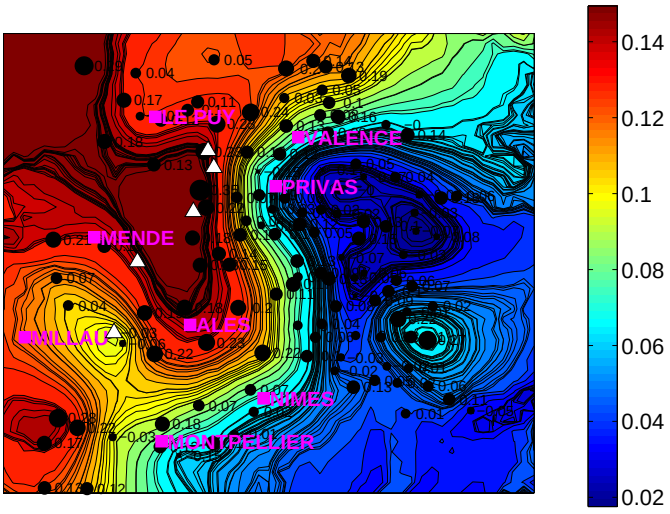

FIG. 3.15 – Estimation par station du paramètre de forme γ par maximum de vraisemblance puis interpolation par krigeage. Pas de temps journalier

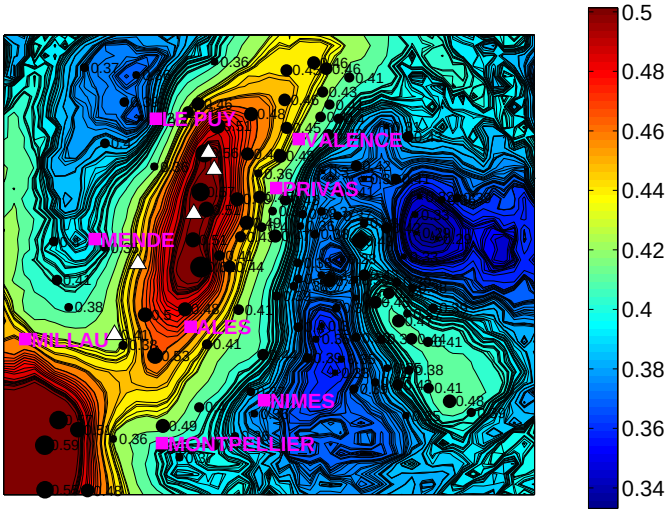

FIG. 3.16 – Estimation par station du paramètre de forme γ par l'estimateur de Hill puis interpolation par krigeage. Pas de temps journalier.

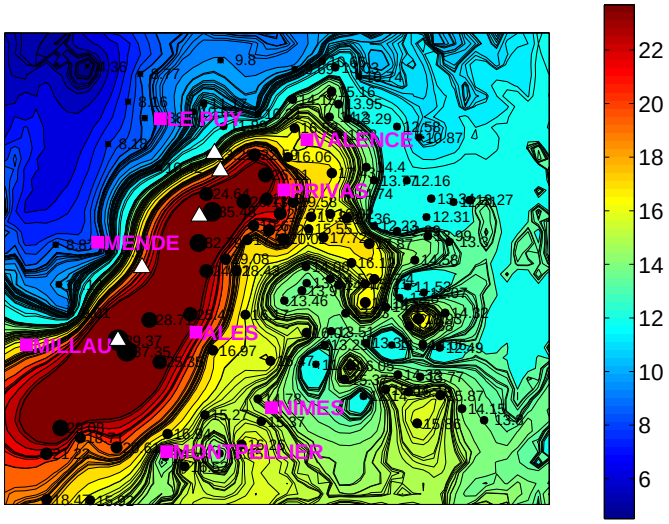

FIG. 3.17 – Estimation par station du paramètre d'échelle σ par maximum de vraisemblance puis interpolation par krigeage. Pas de temps journalier.

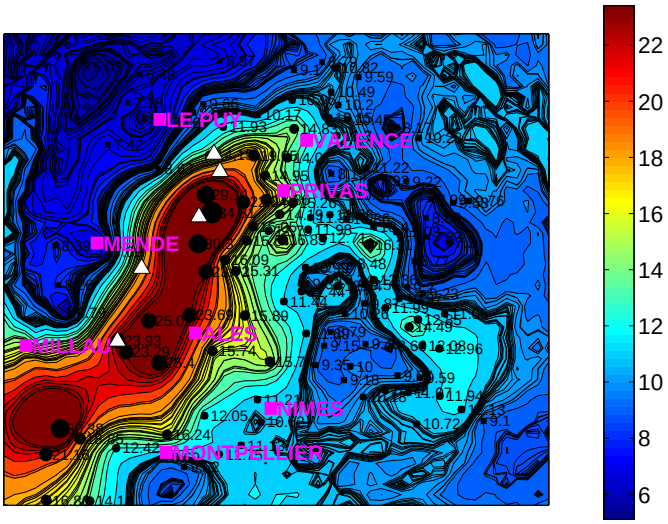

FIG. 3.18 – Estimation par station du paramètre d'échelle σ dans le cas particulier du domaine d'attraction de Fréchet pour lequel $\sigma = u \times \gamma$ où u correspond au seuil et γ au paramètre de forme. Le paramètre de forme γ est estimé par l'estimateur de Hill. Pas de temps journalier.

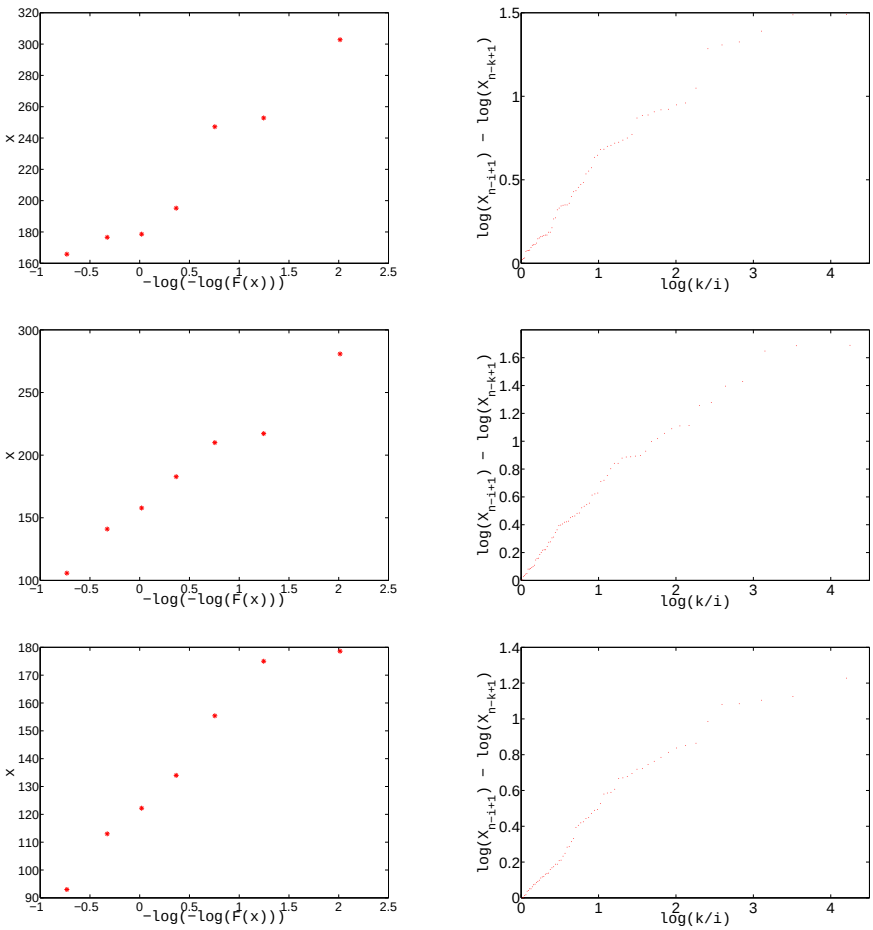

FIG. 3.19 – A gauche : Return level plot. A droite : Diagramme de Hill. Stations de mesures étudiées : Barnas (en haut), Marzan-L'Abbaye (au milieu) et Valleraugue (en bas). Pas de temps journalier.

3.2.2 Cartes des niveaux de retour

Les cartes d'estimation des niveaux de retour pour une période de retour fixée à 10 ans sont très similaires quelle que soit la méthode d'estimation utilisée : Maximum de vraisemblance ou estimateur de Hill. Pour des données horaires, les niveaux de retour sont liés à l'altitude avec des intensités élevées en plaine et plus faibles en montagne (figures 3.20 et 3.21). Pour un cumul journalier des hauteurs de pluie, on observe à l'inverse des valeurs fortes sur la ligne de crêtes, donc plutôt en altitude (figures 3.22 et 3.23). L'altitude ne suffit cependant pas à expliquer les niveaux de retour puisque dans les Alpes, ces derniers sont plutôt faibles, comme en plaine. Globalement, les cartes obtenues pour les niveaux de retour se comportent identiquement à celles obtenues pour le paramètre d'échelle.
On remarque que par l'estimateur de Hill, les niveaux de retour estimés sont plus forts que ceux obtenus par maximum de vraisemblance. La différence est très importante pour les données journalières mais ce résultat s'explique par le fait qu'on a retenu 30% d'excès pour l'approche par maximum de vraisemblance et seulement 10% d'excès pour l'approche Hill. Dans le premier cas, on a donc introduit beaucoup de valeurs 'non extrêmes' qui font chuter les niveaux de retour.

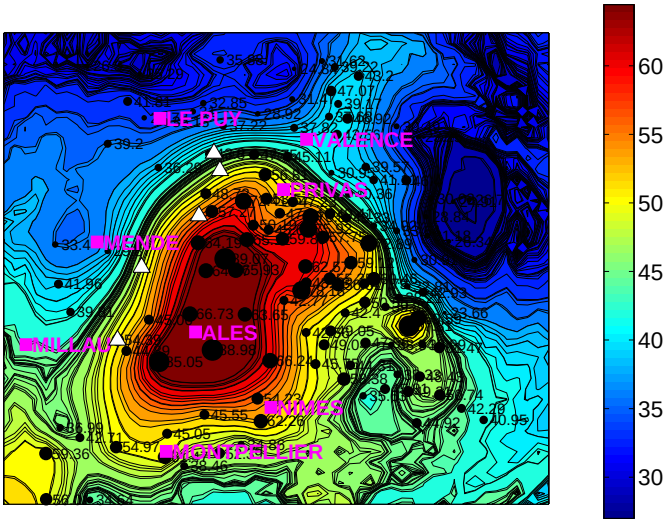

FIG. 3.20 – Niveaux de retour pour une période de retour fixée à 10 ans. Estimation des paramètre de forme et d'échelle par maximum de vraisemblance. Pas de temps horaire. Seuil : 7%.

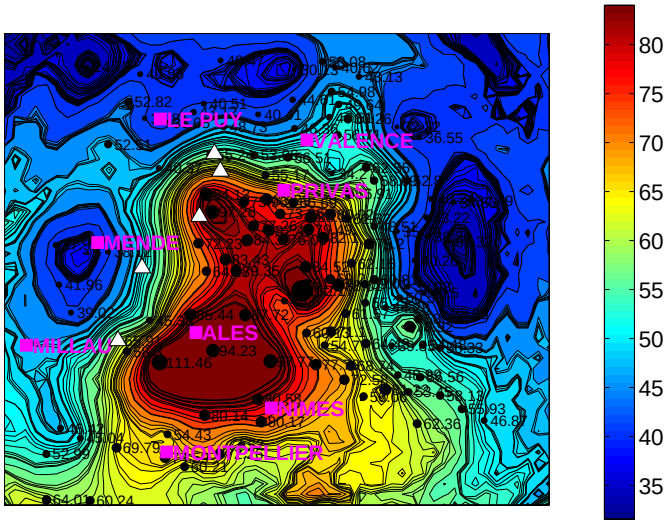

FIG. 3.21 – Niveaux de retour pour une période de retour fixée à 10 ans. On suppose que la fonction de répartition des valeurs extrêmes est dans le domaine de Fréchet. Estimation du paramètre de forme par l'estimateur de Hill. Pas de temps horaire. Seuil : 7%.

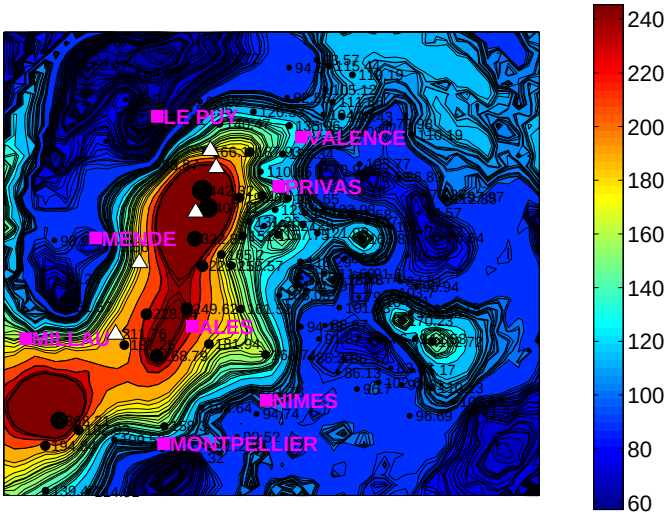

FIG. 3.22 – Niveaux de retour pour une période de retour fixée à 10 ans. Estimation des paramètre de forme et d'échelle par maximum de vraisemblance. Pas de temps journalier. Seuil : 30%.

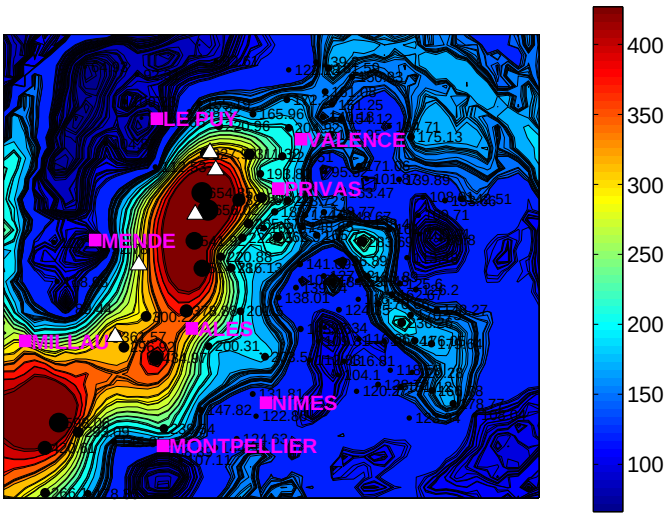

FIG. 3.23 – Niveaux de retour pour une période de retour fixée à 10 ans. On suppose que la fonction de répartition des valeurs extrêmes est dans le domaine de Fréchet. Estimation du paramètre de forme par l'estimateur de Hill. Pas de temps journalier. Seuil : 10%.

3.2.3 Cartes des temps de retour

Les conclusions pour les cartes des temps de retour sont identiques à celles des niveaux de retour. Les niveaux de retour élevés sont ici remplacés par des temps de retour faibles et réciproquement.

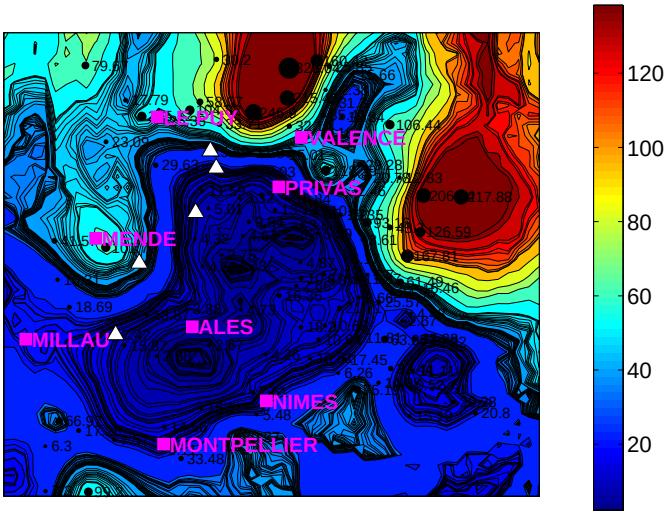

FIG. 3.24 – Temps de retour pour une intensité de pluie fixée à 50mm / heure. Estimation des paramètre de forme et d'échelle par maximum de vraisemblance. Pas de temps horaire. Seuil : 7%.

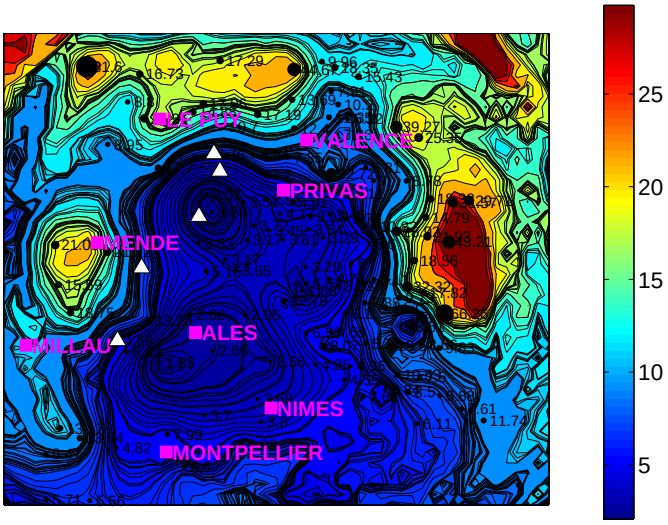

FIG. 3.25 – Temps de retour pour une intensité de pluie fixée à 50mm / heure. On suppose que la fonction de répartition des valeurs extrêmes est dans le domaine de Fréchet. Estimation du paramètre de forme par l'estimateur de Hill. Pas de temps horaire. Seuil : 7%.

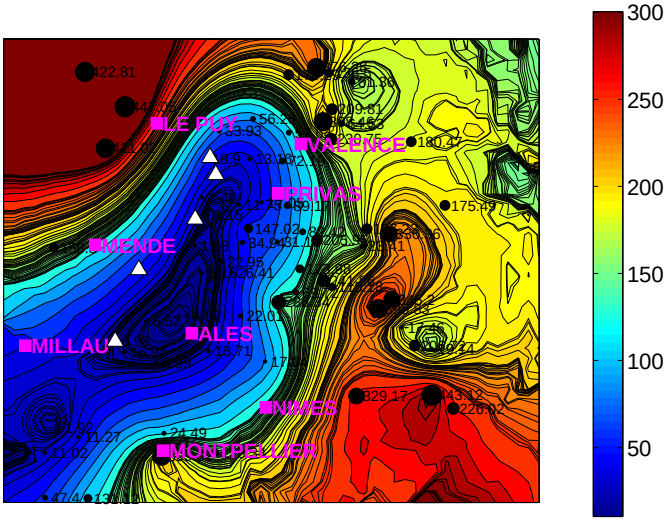

FIG. 3.26 – Temps de retour pour une intensité de pluie fixée à 200mm / jour. Estimation des paramètre de forme et d'échelle par maximum de vraisemblance. Pas de temps journalier. Seuil : 30%.

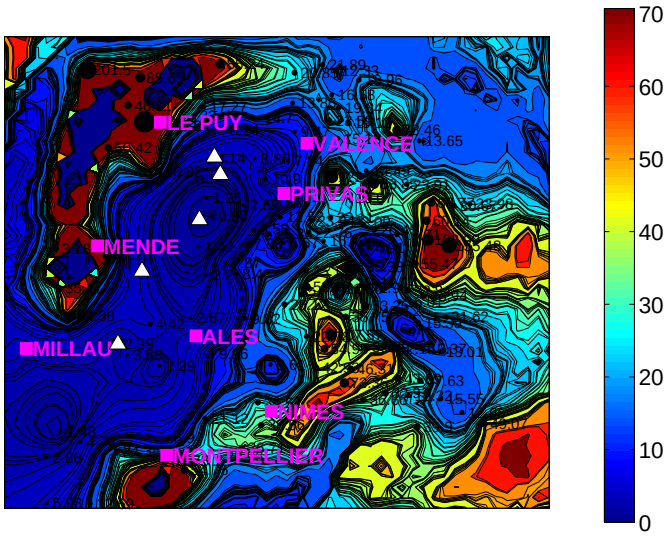

FIG. 3.27 – Temps de retour pour une intensité de pluie fixée à 200mm / jour. On suppose que la fonction de répartition des valeurs extrêmes est dans le domaine de Fréchet. Estimation du paramètre de forme par l'estimateur de Hill. Pas de temps journalier. Seuil : 10%.

3.2.4 Conclusion

En conclusion, cette première analyse des hauteurs de pluie dans la région Cévennes-Vivarais nous a permis de mettre en évidence que pour des données horaires ou journalières, la loi de Gumbel habituellement considérée par les hydrologues ne semble pas la plus pertinente. Le domaine d'attraction de Fréchet est en effet plus adapté pour modéliser la fonction de répartition des valeurs fortes. Se placer dans le domaine de Fréchet a pour avantage de n'avoir qu'un seul paramètre à estimer (le paramètre de forme) et donc une incertitude plus faible sur les estimations des temps et niveaux de retour.
L'analyse des cartes de temps et niveaux de retour donne des conclusions très similaires à celles obtenues par les hydrologues pour les données événementielles (1972-1992) par l'approche "Gumbel + krigeage" : pour des données horaires, les niveaux de retour sont forts en basse altitude alors que pour des données journalières, ils sont forts sur la ligne de crêtes du Massif Central. Inversement, les temps de retour sont faibles en basse altitude pour des données horaires et forts sur la ligne de crêtes pour des données journalières. Les estimations obtenues pour les niveaux de retour par maximum de vraisemblance sont assez proches de celles obtenues dans les études précédentes [3]. Elles sont en revanche plus élevées par l'approche Hill.
Notons que le passage des données horaires aux données journalières pose les difficultés suivantes :

- cumuler les hauteurs de pluie par jour n'est pas pertinent car on divise arbitrairement le temps en blocs de 24h et on ne prend pas en compte les pluies qui s'étalent entre deux jours. D'autre part, on obtient ainsi 24 fois moins de mesures, ce qui rend parfois l'étude des valeurs extrêmes difficile.
- cumuler les hauteurs de pluie sur 24 heures à l'aide d'une fenêtre glissante, comme le font souvent les hydrologues, n'est pas pertinent non plus car on crée ainsi un échantillon de mesures très fortement corrélées dans lequel on a finalement pris en compte 24 fois la même information.

Ces problèmes montrent l'interêt de développer un modèle temporel permettant d'étudier de manière pertinente l'effet des changements d'échelle.
Par ailleurs, dans ce chapitre nous avons supposé que les données étaient indépendantes et identiquement distribuées. Nous verrons dans le chapitre suivant que cette hypothèse est fausse, et qu'un modèle prenant en compte la saisonnalié des mesures et éventuellement la corrélation temporelle doit être développé.
Enfin, les cartographies obtenues dans ce chapitre sont le résultat de deux étapes : une étape de modélisation de la queue de distribution des hauteurs de pluie par station, puis une étape d'interpolation spatiale. Il est alors impossible d'en déduire l'incertitude sur les estimations finales. Un modèle spatial (voire spatio-temporel) devrait donc être développé.

Chapitre 4

Analyse temporelle

Dans ce chapitre, nous présentons une rapide analyse chronologique des données de pluie. Elle met en évidence :

- l'absence de tendance dans les séries,
- la présence d'un effet saisonnier avec des hauteurs de pluie élevées en été-automne et des valeurs plus faibles en hiver-printemps,
- la présence de corrélation temporelle sur une échelle de temps inférieure à la journée.

Ces résultats montrent que les hypothèses d'indépendance et de même loi pour les données, considérées dans le chapitre précédent, ne sont pas vérifiées. Il est donc nécessaire de développer un modèle qui prenne en compte saisonnalité et corrélation temporelle. Dans ce chapitre, nous nous concentrerons sur le développement d'un modèle dans lequel seule la non stationnarité temporelle des pluies est prise en compte. Pour ce faire, nous nous proposons de rechercher un découpage de la série temporelle en saisons homogènes. Le découpage est réalisé par une approche non paramétrique basée sur la statistique du test de Kruskal-Wallis. Une fois le découpage optimal déterminé, les valeurs les plus intenses de pluies sont modélisées par un mélange de loi GPDs dont chaque composante correspondra à une saison. L'ébauche d'un modèle permettant plus de souplesse dans le choix des saisons est présenté.

4.1 Analyse temporelle : tendance, saisonnalité, corrélation ?

L'analyse des séries chronologiques montre que les hypothèses d'indépendance et de même loi pour les hauteurs de pluies $X_1, \ldots, X_n$ ne sont pas vérifiées. La majorité des chroniques présentent une saisonnalité marquée, avec des valeurs fortes en été-automne et des valeurs plus faibles en hiver-printemps, ce qui contredit l'hypothèse d'une même loi pour les hauteurs de pluie au cours du temps. Aucune tendance (croissance ou décroissance systématique des hauteurs de pluie au cours des années) n'est par contre décelée, ce qui n'est pas surprenant sur seulement 8 années.

L'analyse des variogrammes temporels met en évidence une légère corrélation temporelle avec une portée inférieure à la journée.

Comme il est impossible de présenter les chroniques et variogrammes temporels

pour les 142 stations, nous illustrons cette section par un unique exemple jugé représentatif, celui de la station de Glandage. Pour cette station, la chronique des maxima mensuels (figure 4.1 A) montre que les hauteurs de pluies restent stables d'une année à l'autre et qu'il n'y a pas pour ces 7 années d'augmentation ou de diminution systématique des hauteurs de pluie. Elle met en évidence la périodicité des mesures (période d'un an) qui semblent plutôt fortes en été-automne et faibles en hiver-printemps, ce que confirme le graphe (figure 4.1 B), qui présente les boites à moustaches pour les mesures groupées mois par mois. Il est d'ailleurs intéressant de noter que pour cette station, les valeurs les plus fortes sont aux mois de juillet-août. Il serait donc peut être judicieux de continuer à enregistrer les mesures en été contrairement à ce qui est préconisé dans les campagne de mesures actuelles (on ne conserve que les mesures d'automne). Le variogramme temporel calculé avec un pas de temps d'un mois (figure 4.1 C) confirme la périodicité des mesures, qui est d'ailleurs encore plus marquée pour le variogramme que pour la chronique. Il confirme aussi l'absence d'une tendance sur les 7 ans et ne met en évidence aucune corrélation temporelle : ce variogramme pourrait être modélisé par un modèle pépitique ajouté à un modèle de type cosinus.
Comme il n'est pas possible de conclure sur la présence de corrélation temporelle à l'échelle de moins d'un mois pour ce variogramme, nous avons calculé le variogramme temporel pour un pas de temps de 5 heures (figure 4.1 D). Ce variogramme s'ajuste par un modèle de type sphérique avec une portée d'environ une journée, ce qui signifie que les mesures sont corrélées si elles sont séparées par moins d'une journée. Il serait éventuellement intéressant de réaliser une étude plus approfondie de la corrélation temporelle pour l'ensemble des stations car les variogrammes à petit pas de temps ne sont pas tous très structurés.
De cette étude, nous retiendrons toutefois que le point prioritaire à traiter est la prise en compte de la saisonnalité des mesures. La prise en compte des corrélations temporelles reste secondaire.

4.2 Prise en compte de la tendance/saisonnalité

Deux approches sont fréquemment utilisées pour prendre en compte la saisonnalité et/ou la tendance des mesures :

- Soit elles sont incorporées dans les paramètres de forme et d'échelle.
- Soit on découpe l'année en saisons supposées homogènes, c'est à dire sur lesquelles l'hypothèse de stationnarité est acceptable. On définit alors un modèle de mélange sur les saisons.

Nous présentons rapidement ces deux approches. Par la suite, l'approche par saisons est privilégiée et nous proposons une méthode non paramétrique basée sur la statistique du test de Kruskal-Wallis pour choisir au mieux les saisons. Cette méthode est testée sur les données réelles afin de voir si il y a une cohérence dans les résultats.

4.2.1 Approche par modélisation des paramètres de la GPD

Soit X_t la hauteur de pluie à l'instant t. On fixe un seuil $u(t)$. On considère alors que les excès $Y_t = X_t - u(t), X_t > u(t)$ suivent une loi GPD($\gamma(t)$,$\sigma(t)$) où

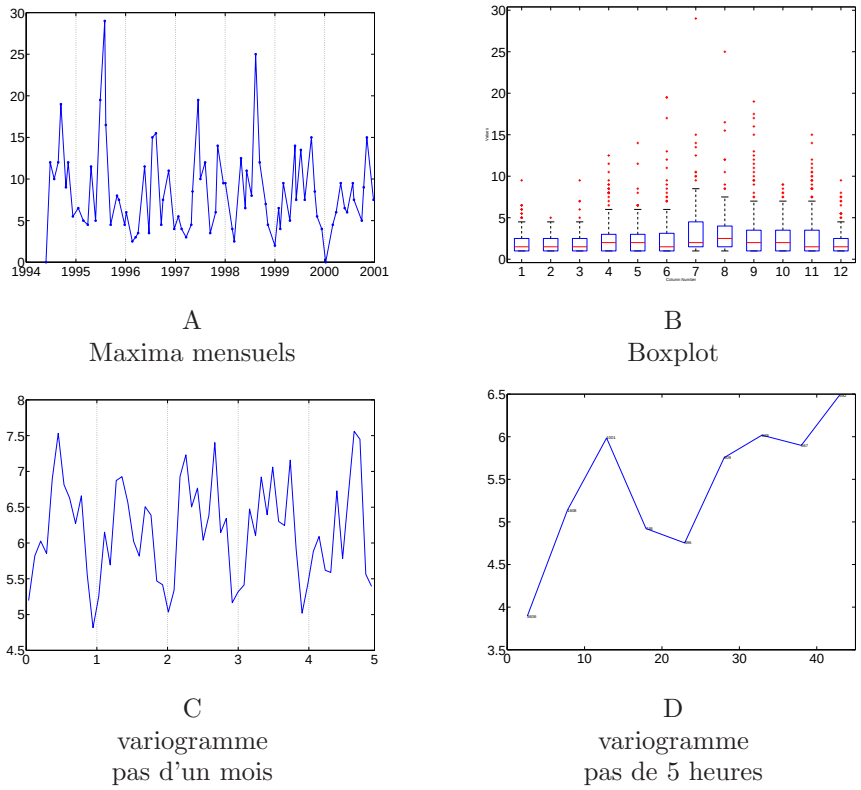

FIG. 4.1 – Etude de la série chronologique de la station de Glandage

$\gamma(t)$ et $\sigma(t)$ sont respectivement les paramètres de forme et d'échelle à l'instant t. Leur fonction de survie $\bar{G}_t$ en y est donc donnée par :

$$\bar{G}_t(y) = (1 + \frac{\gamma(t)}{\sigma(t)} y)^{-\frac{1}{\gamma(t)}}$$

On propose alors de modéliser les paramètres de forme et d'échelle pour prendre en compte la présence d'une tendance ou/et d'une saisonnalité. Par exemple pour étudier les hauteurs de pluies, Coles propose un paramètre de forme constant et une tendance linéaire pour le logarithme du paramètre d'échelle ([8], chapitre 6, p119) :

$$\sigma(t) = \exp(\beta_0 + \beta_1 t)$$

Dans ce cas, la périodicité des mesures n'est pas prise en compte ce qui peut être fait en ajoutant un modèle de type cosinus :

$$\sigma(t) = \exp(\beta_0 + \beta_1 t + \beta_2 (1 - \cos \frac{2\pi t}{T}))$$

où T est la période.
Une fois les modèles choisis pour les paramètres de forme et d'échelle, les paramètres de ces modèles sont estimés par maximum de vraisemblance.
Cette approche a plusieurs avantages :

– Elle est facile à implémenter.
– Elle propose un modèle temporel avec lequel on peut prédire en fonction du temps la probabilité de dépasser un seuil.

Ses inconvénients sont les suivants :

– Le nombre de paramètres à estimer croît avec la complexité du modèle choisi pour les paramètres de forme et d'échelle. Or on a pu voir dans le chapitre précédent qu'avec peu de mesures, les estimations par maximum de vraisemblance ne sont pas toujours très fiables et sont entachées d'une erreur importante. En augmentant le nombre de paramètres à estimer, on risque d'augmenter fortement les incertitudes sur les estimations.
– Le choix du modèle est difficile. Sur simulations, lorsqu'on choisit un modèle pour $\gamma(t)$ et $\sigma(t)$, cette approche marche parfaitement bien, mais sur des données concrètes, il est parfois difficile de choisir un modèle pour les paramètres. Prenons l'exemple des trois stations de Barnas, Mazan l'Abbaye et Valleraugue. L'estimation par maximum de vraisemblance (figure 4.2) et par Hill (figure 4.3) des paramètres de forme et d'échelle par mois montre que le choix d'un modèle est difficile. Le paramètre d'échelle prend des valeurs fortes en été-automne et faibles au printemps/hiver, mais il n'est pas évident d'ajuster un modèle. Par ailleurs, le comportement des paramètres diffère d'une station à une autre et il est impossible de proposer un seul modèle pour toutes les stations. Une étude au cas par cas est donc nécessaire.
– Enfin, le choix d'un seuil en fonction du temps n'est pas non plus simple.

Pour une analyse plus approfondie de ces méthodes et de leur application, le lecteur peut se reporter à [8, 13, 18, 24, 23, 5].

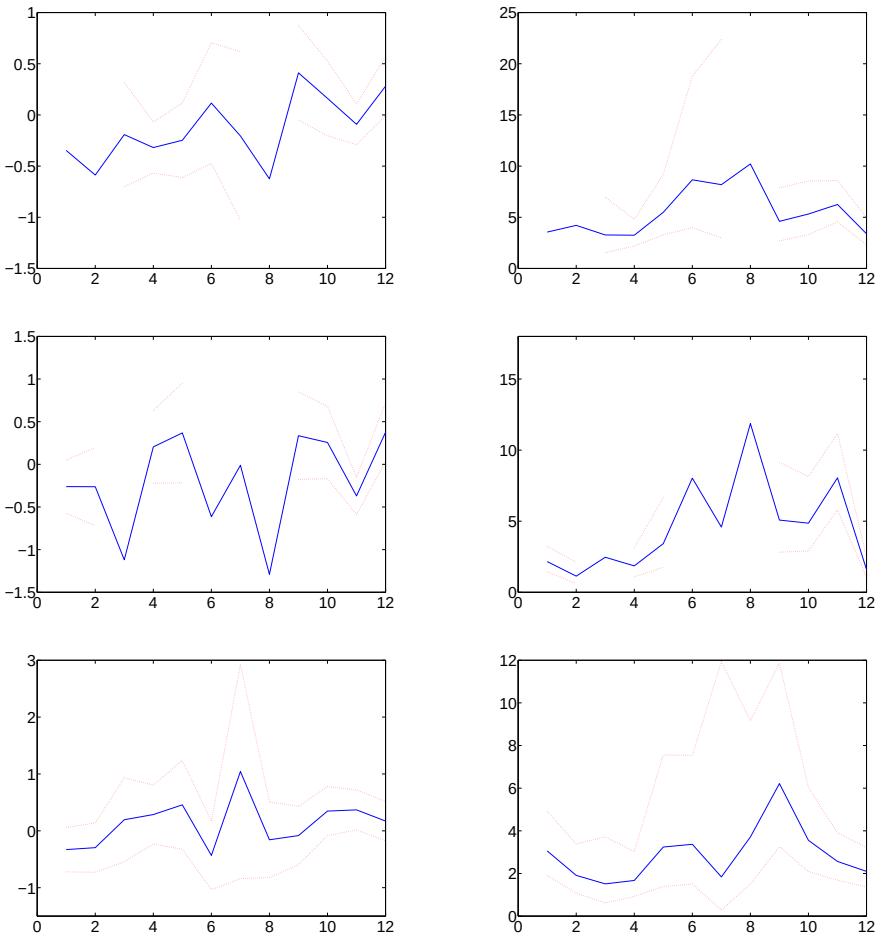

FIG. 4.2 – Estimation par mois du paramètre de forme (à gauche) et du paramètre d'échelle (à droite) pour les stations de Barnas, Mazan L'Abbaye et Valleraugue (de haut en bas). Pour chaque station, toutes les mesures ont été groupées par mois sans tenir compte de l'année. Nous avons ensuite supposé que les excès par mois suivent une loi GPD dont les paramètres ont été estimés par maximum de vraisemblance. Le pourcentage d'excès par mois a été fixé à 10%.

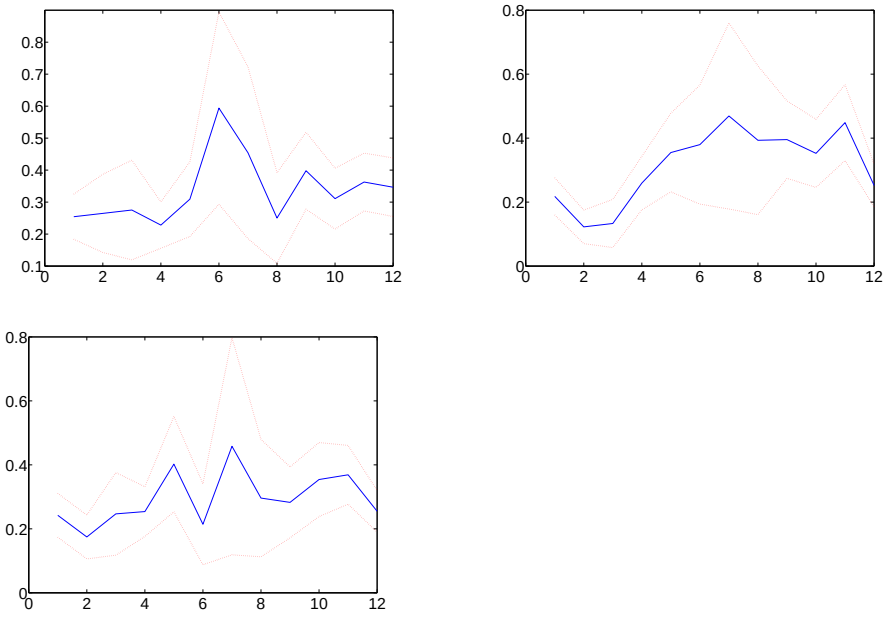

FIG. 4.3 – Estimation par mois du paramètre de forme (à gauche) pour les stations de Barnas, Mazan L'Abbaye et Valleraugue (de haut en bas). Pour chaque station, toutes les mesures ont été groupées par mois sans tenir compte de l'année. Nous avons ensuite supposé que les excès par mois suivent une loi GPD dont les paramètres ont été estimés par l'estimateur de Hill. Le pourcentage d'excès par mois a été fixé à 10%.

4.2.2 Approche par saisons

Une autre approche consiste à découper la chronique temporelle en blocs sur lesquels l'hypothèse de stationnarité est acceptable. On l'appelle l'approche par saisons. Elle est utilisée dans [21, 27, 9, 10, 13, 27]. Lorsqu'il n'y a pas de tendance dans la chronique, on peut grouper les années entre elles, ce qui revient à ne pas prendre en compte l'année de mesure mais uniquement le jour et le mois de la mesure. C'est ce que nous ferons tout au long de ce rapport suite aux conclusions de la section 4.1 dans laquelle nous avons montré qu'il n'y a pas de tendance dans les séries chronologiques de hauteurs de pluie.
L'idée est donc de découper l'année en R saisons jugées homogènes pour les valeurs fortes. On note $S_1, \ldots, S_R$ les R saisons. Pour chaque saison on définit un seuil $u_j, j \in \{1, \ldots, R\}$. On suppose que dans chaque saison j, les excès $Y = (X - u_j)1_{\{X \in S_j, X > u_j\}}$ suivent une loi GPD(γ_j, σ_j). La loi de X sur l'année est définie par un modèle de mélange dont la densité g est donnée par :

$$g(x) = \sum_{i=1}^{S} p(X = x|S_i) \times p(X \in S_i)$$

où $p(X = x|S_i)$ est la densité de X sachant qu'on est dans la saison S_i et $p(X \in S_i)$ est la probabilité de pluie dans la saison S_i. La loi des excès étant connue, on en déduit facilement la probabilité que la hauteur de pluie X dépasse un seuil x, si x est suffisamment grand :

$$\begin{aligned}\mathbb{P}(X > x) &= \sum_{i=1}^{R} \mathbb{P}(X > x|X \in S_i)\mathbb{P}(X \in S_i) \\ &= \sum_{i=1}^{R} \mathbb{P}(X > x|X > u_i, X_i \in S_i) \times \mathbb{P}(X > u_i|X \in S_i) \times \mathbb{P}(X \in S_i)\end{aligned}$$

Cette quantité est estimée par :

$$\hat{\mathbb{P}}(X > x) = \sum_{i=1}^{R} (1 + \gamma_i \frac{x - u_i}{\sigma_i})^{-\frac{1}{\gamma_i}} \times \frac{n_i}{N_i} \times \frac{N_i}{n} \tag{4.1}$$

$$= \sum_{i=1}^{R} (1 + \gamma_i \frac{x - u_i}{\sigma_i})^{-\frac{1}{\gamma_i}} \times \frac{n_i}{n} \tag{4.2}$$

où n_i est le nombre d'excès dans la saison i, N_i est le nombre de mesures dans la saison i et n le nombre de mesures sur l'année. La probabilité $\mathbb{P}(X > x)$ s'interprète de la manière suivante : c'est la probabilité que la hauteur de pluie, à l'instant t tiré aléatoirement sur l'année, soit supérieure à x.

Notons qu'en revanche, le calcul des niveaux de retour associés à une période de retour n'est pas explicite.
Souvent, le choix des saisons est guidé par l'expérience et la connaissance des hydrologues. Il n'est cependant jamais évident de dire avec certitude le nombre de saisons à retenir et quels mois affecter à ces saisons. C'est pourquoi nous proposons dans la section suivante une méthode non paramétrique, basée sur la statistique du test de Kruskal-Wallis, pour choisir les saisons.

Choix des saisons

Dans cette section, on cherche à diviser l'année en R saisons continues, c'est à dire qu'une saison doit forcément contenir des mois consécutifs. Notons qu'il existe en fait C_{12}^R manières de découper l'année en R saisons. Effectivement, en représentant l'année sous la forme d'un disque divisé en 12 parties égales (voir figure 4.4), on voit aisément que choisir R saisons revient à choisir R points parmi les 12 points délimitant les parties du disque. Il y a donc C_{12}^R manières de découper une année en R saisons.

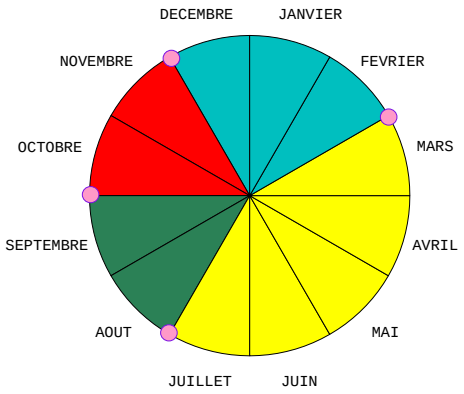

FIG. 4.4 – Représentation d'une année sous forme d'un disque

L'idée de notre approche est de se fixer un nombre R de saisons et de parcourir l'ensemble des C_{12}^R découpages possibles. On retiendra alors le découpage le meilleur selon un critère prédéfini. Deux critères non paramétriques basés sur la statistique du test de Kruskal-Wallis sont étudiés ici :

1. Critère Kruskal : Pour chaque découpage en saisons, on calcule la statistique du test de Kruskal-Walis qui compare le rang moyen des observations dans chaque saison au rang moyen sur l'année. Cette statistique sera élevée lorsque les saisons sont bien démarquées, c'est à dire avec des valeurs très différentes. Elle sera au contraire faible si les saisons se ressemblent. L'idée est alors de maximiser la statistique de Kruskal-Wallis sur les C_{12}^R découpages possibles et de retenir le découpage en saisons qui sépare au mieux les saisons.
   Description de la méthode :
   On découpe l'année en R saisons. Soient :
   – N_i le nombre d'observations dans la saison i
   – n le nombre total d'observations sur l'année
   On considère l'ensemble des mesures sur l'année et on calcule le rang de chacune des mesures. On note alors :
   – r_{ij} le rang de l'observation j de la saison i

– $\bar{r}_i$ la moyenne des rangs des observations de la saison i

$$\bar{r}_i = \frac{\sum_{j=1}^{N_j} r_{ij}}{N_i}$$

– $\bar{r}$ la moyenne de tous les r_{ij}

$$\bar{r} = \frac{n+1}{2}$$

La statistique du test de Kruskal-Wallis est donnée par :

$$K = (N-1)\frac{\sum_{i=1}^{R} N_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{R}\sum_{j=1}^{N_i}(r_{ij} - \bar{r})^2}$$

Notons que le dénominateur de K vaut exactement $(N-1)N(N+1)/12$. D'où

$$K = \frac{12}{N(N+1)}\sum_{i=1}^{R} N_i(\bar{r}_i - \bar{r})^2$$

On cherche alors le découpage en saisons qui maximise K, c'est à dire qui maximise $K' = \sum_{i=1}^{R} N_i(\bar{r}_i - \bar{r})^2$. Par la suite, nous appellerons K' la statistique de "Kruskal-Wallis simplifiée".

2. Critère P-valeur : Après avoir découpé l'année en R saisons, on effectue pour chacune des saisons, un test de Kruskal-Wallis entre les mois de la saison. Ce test permet de déterminer si les mois dans la saison suivent la même loi. Une p-valeur élevée (supérieure à au moins 0.05) indique les mesures des mois de la saison suivent bien la même loi. A l'inverse une p-valeur faible signifie que la loi n'est pas la même pour tous les mois de la saison et que le découpage en saisons est donc probablement mauvais.
   Description de la méthode :
   – On découpe l'année en R saisons.
   – On calcule la p-valeur P_i du test de Kruskal-Wallis pour chaque saison S_i.
   – On calcule alors le coût $C = 1 - \min_{i\in 1,\ldots,R}(P_i)$.
   – On minimise les coûts sur l'ensemble des C_{12}^R découpages possibles.

Afin de tester ces deux méthodes, nous avons simulé des mesures de pluie sur une année, pour lesquelles les saisons sont bien démarquées. La loi de Student a été retenue pour les simulations car elle fournit des jeux de données assez comparables aux observations. Pour chaque mois, 1000 mesures de pluie ont été simulées selon une loi de Student dont le paramètre dépend de la saison auquel le mois appartient. Ces saisons ont été choisies de la manière suivante :

– Saison 1 : janvier, février, mars
– Saison 2 : avril, mai, juin, avril
– Saison 3 : août, septembre
– Saison 4 : octobre, novembre, décembre

Notons que si nous décrivons les $C_{12}^4 = 495$ découpages possibles, ces saisons correspondent au 101ème découpage. Deux cas ont été envisagés pour les simulations :

– Simulation 1 : les saisons sont bien démarquées en moyenne et en variance (figure 4.5 A)

– Simulation 2 : les saisons sont démarquées uniquement en variance (la moyenne et la médiane sont les mêmes pour tous les mois, voir figure 4.5 B)

Les valeurs des paramètres de la loi de Student pour les deux simulations sont présentées table 4.1.

Mois	Simulation 1	Simulation 2
Janvier	2+ $\lvert Student(3)\rvert$	$\lvert Student(3)\rvert$
Février	2+$\lvert Student(3)\rvert$	$\lvert Student(3)\rvert$
Mars	2+$\lvert Student(3)\rvert$	$\lvert Student(3)\rvert$
Avril	10+$\lvert Student(2)\rvert$	$\lvert Student(2)\rvert$
Mai	10+$\lvert Student(2)\rvert$	$\lvert Student(2)\rvert$
Juin	10+$\lvert Student(2)\rvert$	$\lvert Student(2)\rvert$
Juillet	10+$\lvert Student(2)\rvert$	$\lvert Student(2)\rvert$
Août	4+$\lvert Student(8)\rvert$	$\lvert Student(8)\rvert$
Septembre	4+$\lvert Student(8)\rvert$	$\lvert Student(8)\rvert$
Octobre	15 +$\lvert Student(5)\rvert$	$\lvert Student(5)\rvert$
Novembre	15 +$\lvert Student(5)\rvert$	$\lvert Student(5)\rvert$
Décembre	15 +$\lvert Student(5)\rvert$	$\lvert Student(5)\rvert$

TAB. 4.1 – Lois simulées par mois pour la simulation 1 et pour la simulation 2.

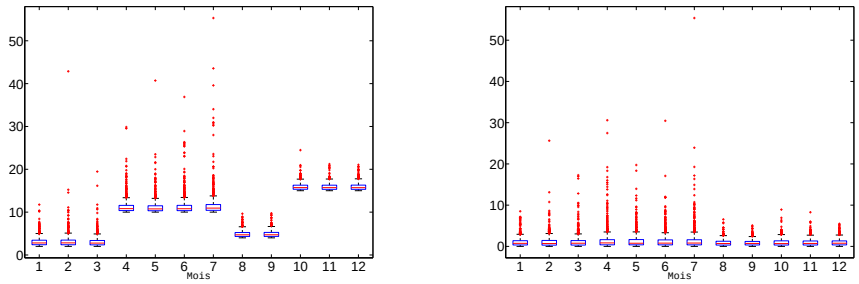

FIG. 4.5 – Données simulées (1000 simulations par mois). A : simulation 1. B : simulation 2.

Les résultats sur le choix des saisons pour chacune des simulations sont présentés figure 4.6 pour l'approche par Kruskal et figure 4.7 pour l'approche par P-valeur.

Simulation 1 :

– Approche Kruskal : La figure 4.6 A présente l'évolution de la statistique de Kruskal-Wallis en fonction du numéro du découpage en saisons. Elle montre en particulier que le maximum est atteint pour le 101ème découpage : La méthode permet donc bien de retrouver les saisons simulées. Les 10 valeurs les plus fortes sont entourées par un cercle. Dans la figure C, on présente les saisons correspondant à ces solutions. Si on regarde les 9 autres solutions trouvées, on voit que les deux premières solutions sont très proches du découpage optimal puisqu'elles s'en écartent

seulement d'un mois. A partir de la 4ème solution, on commence à s'éloigner de plus en plus de la vraie solution.
La figure E montre l'évolution de la statistique de Kruskal-Wallis simplifiée (la statistique maximale sur l'ensemble des découpages) en fonction du nombre de saisons retenues pour l'étude. Dans notre exemple, on sait que 4 saisons ont été simulées, il est donc judicieux de rechercher 4 saisons. Pour des données réelles, on ne sait pas combien de saisons il faudra retenir. Nous avons donc regardé quelles saisons étaient trouvées par la méthode si l'année est découpée en $2, 3, \ldots, 11$ saisons. Les saisons obtenues sont présentées figure G et la statistique de Kruskal-Wallis associée à ces saisons en fonction de la taille du découpage est présentée figure E. On voit que les saisons trouvées sont tout à fait cohérentes entre elles : la solution trouvée pour un découpage en $R+1$ saisons correspond à la solution trouvée pour un découpage en R saisons dans laquelle une des saisons est découpée en deux. Cela explique la croissance de la statistique de Kruskal sur la figure E et le fait qu'elle ne croît quasi plus à partir d'un découpage en plus de 4 saisons. Effectivement, pour moins de 4 saisons, on groupe forcément dans une même saison des mois qui ne sont pas de même loi. Pour un découpage en 4 saisons, on trouve les bonnes saisons, donc la statistique de Kruskal est nettement plus élevée qu'avec un découpage en 2 et 3 saisons. Pour plus de 4 saisons, on redivise des saisons qui étaient bien trouvées, on ne groupe donc pas des mois de lois différentes, on ne fait que rediviser des saisons en sous-saisons, la statistique de Kruskal reste donc constante. Ce graphique est intéressant dans la mesure où il nous permet d'avoir une idée du nombre de saisons à retenir.

– Approche P-valeur : La figure 4.7 A montre que la fonction de coût atteint son minimum pour le 101ème découpage de saison : les saisons simulées sont donc bien retrouvées. La méthode est radicale sur le choix du découpage puisque la fonction de coût est égale à 1 pour tous les découpages sauf pour le découpage simulé. L'inconvénient de cette méthode est qu'on ne peut pas comparer les fonctions de coût entre elles pour différents découpage en saisons. On a donc aucune intuition sur le nombre de saisons à retenir. Par ailleurs, cette approche est plus compliquée puisqu'on doit effectuer R tests de Kruskall-Wallis par découpage en R saisons. Dans la suite, nous abandonnerons donc cette approche et privilégierons l'approche "Kruskal".

Simulation 2 :

Pour la simulation 2, les résultats sont présentés figure 4.6 et figure 4.7 dans la colonne droite. Aucune des deux méthodes ne permet de retrouver les saisons simulées. Ce résultat n'est pas particulièrement étonnant car il est difficile de séparer les groupes lorsque les médianes sont égales. Or, l'analyse des données expérimentales montre justement que les médianes sont relativement proches entre les mois. En revanche, la variance des mesures par mois diffère considérablement d'un mois à l'autre.
En fait, il n'est pas judicieux de travailler sur la totalité des mesures étant donné que ce sont uniquement les valeurs fortes qui nous intéressent. De plus, en conservant uniquement les valeurs fortes, par exemple en conservant 10% de dépassements par mois, on amplifie la non-stationnarité des mesures et on sépare plus facilement les saisons. Deux approches sont possibles pour choisir les valeurs fortes sur lesquelles travailler :

– Approche seuil par saisons : On se fixe un seuil par saison. Pour chaque découpage testé, on conserve 10% des dépassements par saison.
– Approche seuil par mois : On se fixe un seuil par mois, par exemple on conserve 10% de dépassements par mois (voir les boites à moustache pour les simulation 1 et 2 figure 4.8). En travaillant sur les dépassements, on voit alors que l'effet saisonnier est plus marqué.

Les résultats pour l'approche seuil par saison sont présentés figure 4.9. On voit que les saisons sont cette fois bien trouvées pour la simulation 1 mais aussi pour la simulation 2. En conservant les 10 plus fortes valeurs de la statistique de Kruskal-Wallis simplifiée, on trouve 10 découpages en saisons relativement cohérents. Par contre, il y a certaines incohérences dans les résultats lorsque l'on fait varier la taille du découpage en saisons. Par exemple pour un découpage en 5 saisons, mars et avril sont groupés en saison alors qu'ils n'ont pas été simulés avec la même loi. Cette erreur est due au fait qu'on conserve 10% des dépassements par saison et non par mois. Comme les valeurs du mois d'avril sont nettement plus élevées que celles de Mars, lorsqu'on retient 10% des dépassements pour la saison [Mars Avril], seules les mesures du mois d'avril vont rester. Pour mieux séparer les saisons, il est en fait préferable de travailler sur des dépassements calculés par mois. Ce que confirment les résultats pour cette approche, présentés figure 4.10. On voit que par l'approche seuil par mois, les saisons sont bien retrouvées pour les deux simulations, les solutions sont plus marquées, et le choix des saisons reste cohérent quelle que soit la taille du découpage.

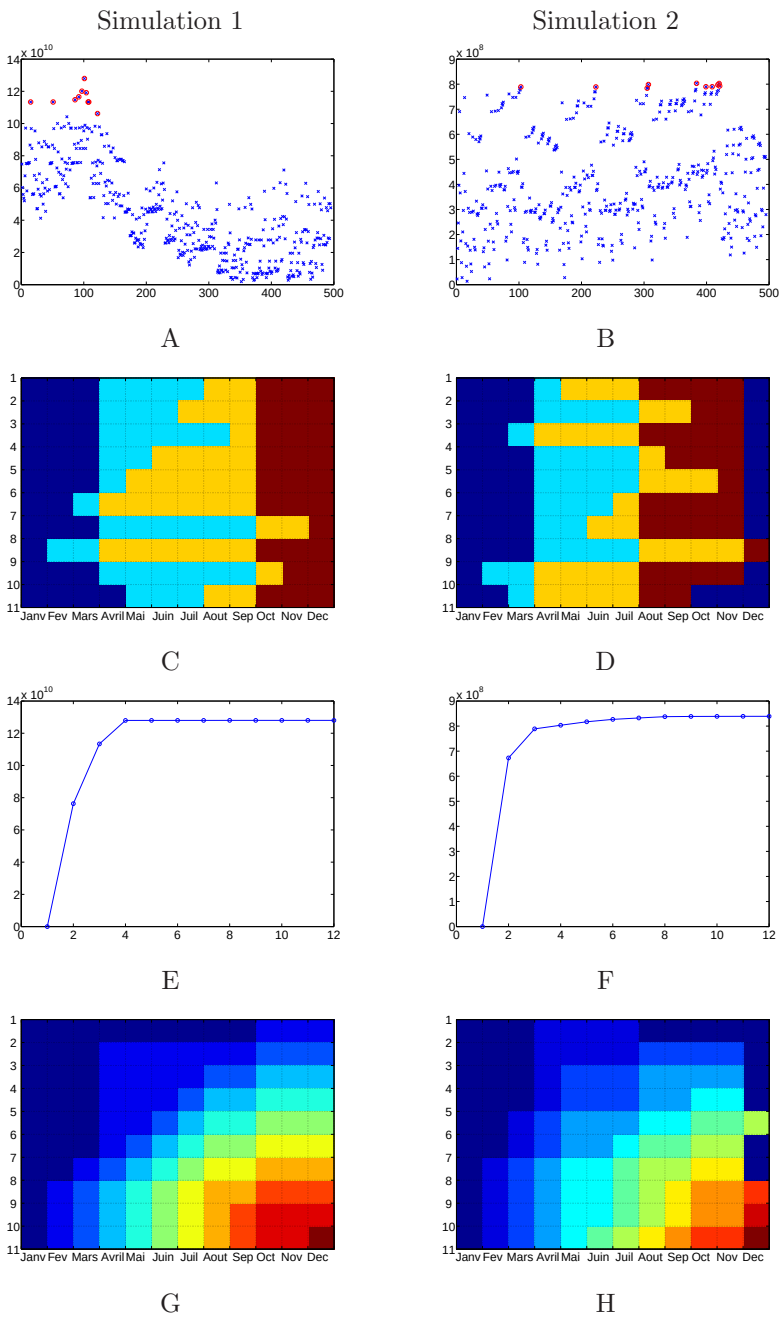

FIG. 4.6 – Choix des saisons sur l'ensemble des mesures. A gauche : Résultats pour la simulation 1. A droite : Résultats pour la simulation 2. A et B : Statistique de Kruskal-Wallis simplifiée en fonction des C_{12}^4 découpages en saisons possibles. C et D : Présentation des 10 meilleures saisons trouvées par la méthode. La meilleure solution se situe en haut du graphe. Les mois d'une même saison sont de la même couleur. E et F : Evolution de la statistique de Kruskal-Wallis simplifiée pour la meilleure saison en fonction du nombre de saisons. G et H : Présentation des saisons trouvées en fonction de la taille du découpage en saisons

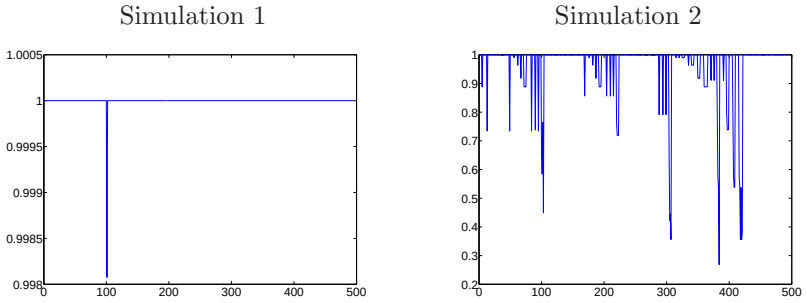

FIG. 4.7 – Abscisses : numéro de la saison (C_{12}^4 configurations de saisons possibles). Ordonnées : fonction de coût pour l'approche par p-valeur.

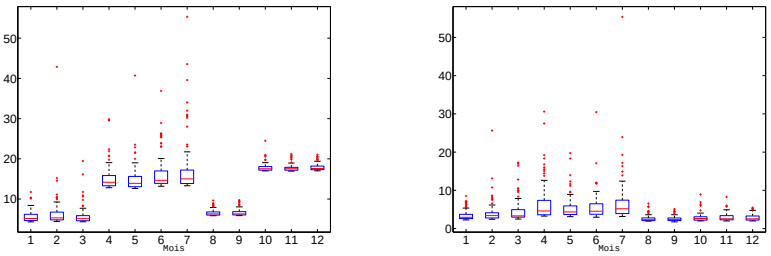

FIG. 4.8 – Boites à moustache par mois en conservant 10% de dépassements par mois. A gauche : simulation 1. A droite : simulation 2.

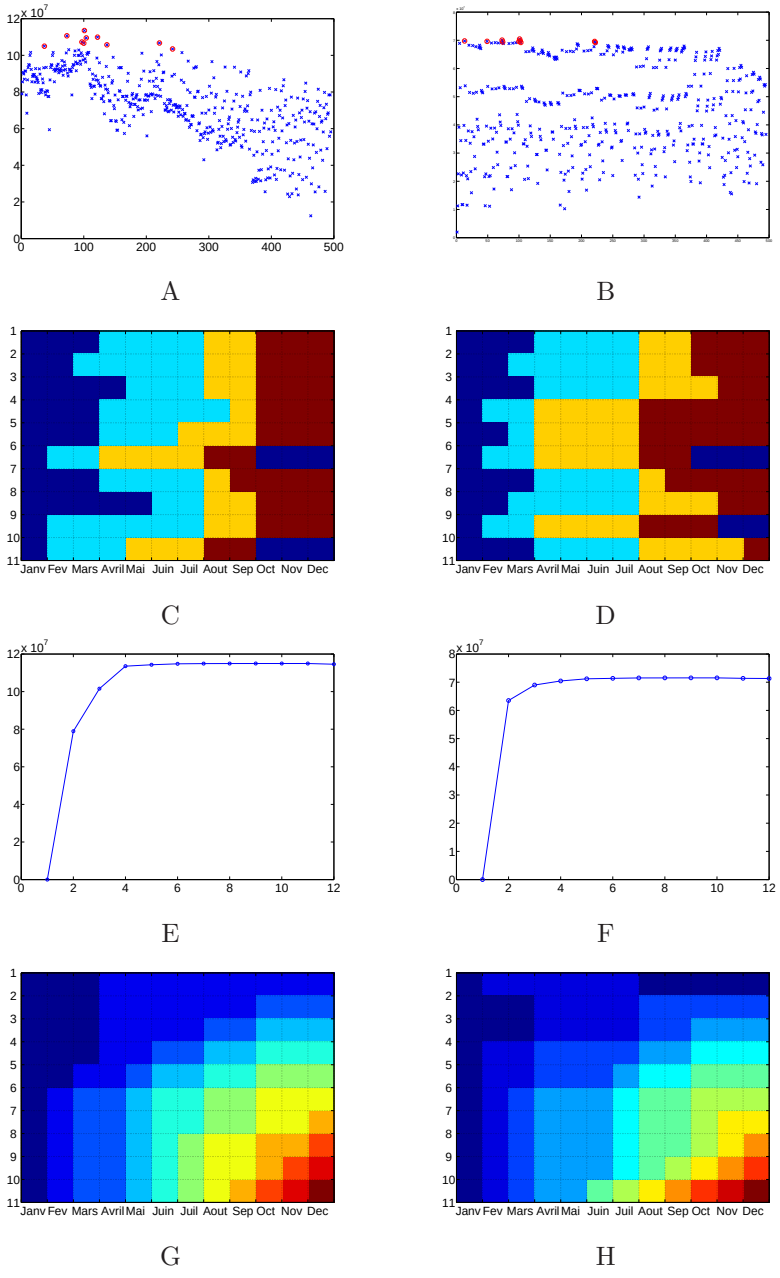

FIG. 4.9 – Choix des saisons. 10% de dépassements par saison. A gauche : Résultats pour la simulation 1. A droite : Résultats pour la simulation 2. A et B : Statistique de Kruskal-Wallis simplifiée en fonction des C_{12}^4 découpages en saisons possibles. C et D : Présentation des 10 meilleures saisons trouvées par la méthode. La meilleure solution se situe en haut du graphe. Les mois d'une même saison sont de la même couleur. E et F : Evolution de la statistique de Kruskal-Wallis simplifiée pour la meilleure saison en fonction du nombre de saisons. G et H : Présentation des saisons trouvées en fonction de la taille du découpage en saisons

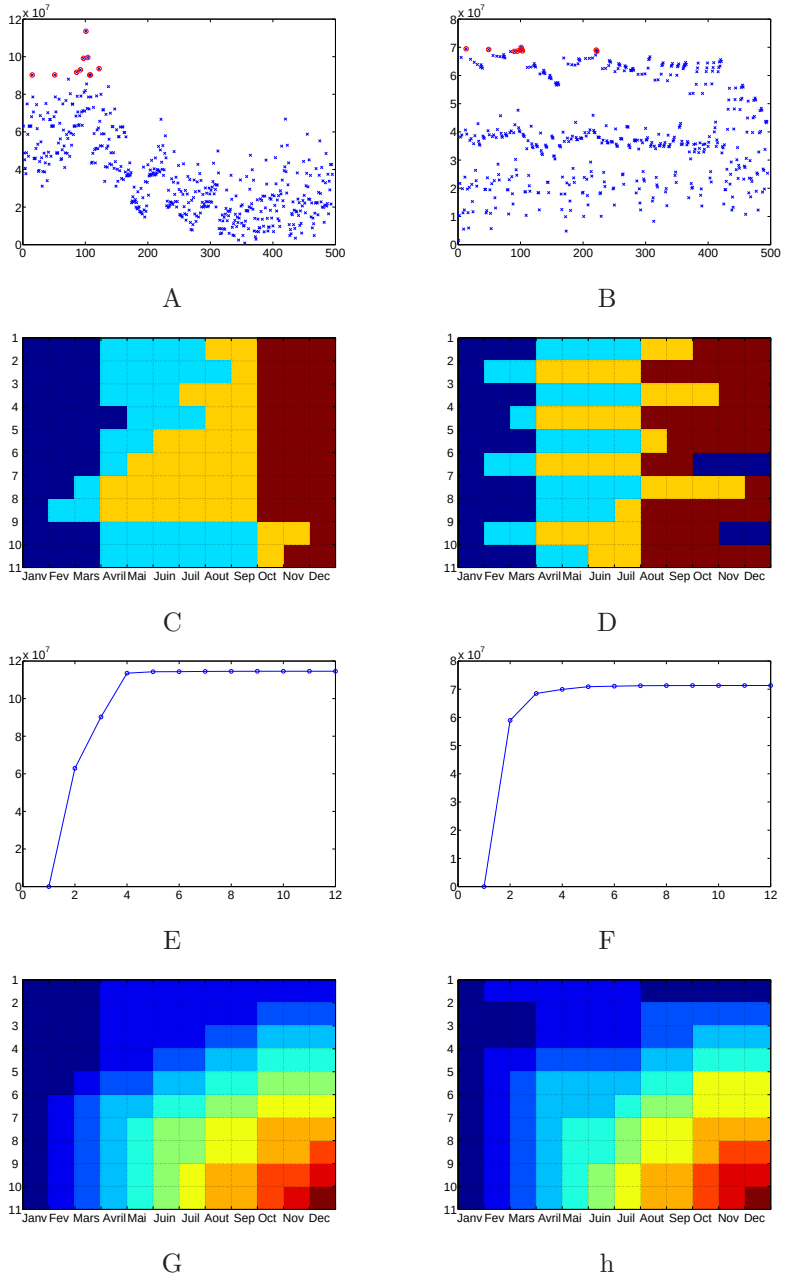

FIG. 4.10 – Choix des saisons. 10% de dépassements par mois. A gauche : Résultats pour la simulation 1. A droite : Résultats pour la simulation 2. A et B : Statistique de Kruskal-Wallis simplifiée en fonction des C_{12}^4 découpages en saisons possibles. C et D : Présentation des 10 meilleures saisons trouvées par la méthode. La meilleure solution se situe en haut du graphe. Les mois d'une même saison sont de la même couleur. E et F : Evolution de la statistique de Kruskal-Wallis simplifiée pour la meilleure saison en fonction du nombre de saisons. G et H : Présentation des saisons trouvées en fonction de la taille du découpage en saisons

Application aux données réelles L'approche Kruskall a été appliquée sur l'ensemble des stations de la région Cévennes-Vivarais afin de découper les chroniques en saisons. Nous avons retenu 10% de dépassements pour effectuer cette étude et nous avons testé les découpages en 2, 3, 4 et 5 saisons pour voir la cohérence entre les résultats. Les résultats sont présentés figure 4.11.
Les résultats trouvés pour un découpage en deux saisons sont assez cohérents pour l'ensemble des stations : sur la figure 4.11 A, on distingue en gros une saison allant de mai à novembre et une saison allant de décembre à avril. Pour un découpage en 3 saisons, les résultats sont plus partagés mais on retrouve une certaine cohérence avec le découpage en deux saisons. Ils donnent finalement l'impression qu'un découpage en deux saisons serait préférable car on continue à avoir des saisons assez similaires à celles obtenues pour le découpage en deux saisons et la troisième saison est généralement constituée par un ou deux mois seulement. Avec 4 saisons, on trouve toujours une certaine cohérence entre les résultats. On distingue en gros une saison de décembre à mars, une saison d'avril à juin, une saison de juillet à septembre et une saison d'octobre à novembre. Avec 5 saisons ou plus, les résultats deviennent de plus en plus difficiles à interpréter.

Pour résumer l'information contenue dans ces figures, il serait intéressant de savoir en moyenne quel est le découpage en saisons à retenir pour l'ensemble des stations. Pour ce faire, nous proposons la méthode suivante. Pour chacune des stations $j \in 1, \ldots, 142$, la statistique de Kruskal simplifiée K'_{ji} a été calculée pour chaque découpage i en R saisons possibles ($i = 1, \ldots, C_{12}^R$). Si l'on moyenne les K'_{ji} sur les j stations, et qu'on maximise par rapport à i cette quantité, on obtient le découpage en R saisons voulu.
Appliqué au mesures, ce calcul propose les découpages suivants :

- Découpage en 2 saisons : [dec janv fev mars avril] et [mai juin juillet août septembre, octobre, novembre]
- Découpage en 3 saisons : [dec janv fev mars], [avril] et [mai juin, juillet août, septembre, octobre, novembre]
- Découpage en 4 saisons : [dec janv fev mars], [avril, mai juin], [juillet août septembre], [octobre novembre]
- Découpage en 5 saisons : [dec janv fev mars], [avril], [mai, juin], [juillet août, septembre] et [octobre, novembre]

Ces découpages en saisons sont relativement cohérents entre eux quel que soit le découpage choisi, sauf pour le mois d'avril, qui se retrouve une fois avec mars, une fois avec mai et deux fois seuls. Par ailleurs, ils semblent bien résumer les découpages présentés figure 4.11. Cependant, on peut voir sur la figure 4.11 que les découpages en saisons diffèrent pour chacune des stations, en particulier quand on augmente le nombre de stations. Nous avons cherché à voir si il y avait une cohérence spatiale dans les découpages trouvés, c'est à dire si pour une station et sa voisine, on trouve bien un découpage en saisons similaire. L'objectif est aussi de voir si on distingue différents régimes de pluie dans la région. Une manière de présenter les résultats serait de représenter par une même couleur toutes les stations pour lesquelles un même découpage en saisons a été choisi. Cependant, sur l'ensemble des stations, on trouve jusqu'à 30 ou 40 découpages différents. Les représenter tous par des couleurs différentes ne semble pas très pertinent à moins que les couleurs soient choisies de telle manière que deux

découpages similaires ont des couleurs similaires. Mais le choix des couleurs est alors difficile car il faut d'abord définir ce qu'on entend par découpage similaire. Nous avons donc plutôt choisi de segmenter les stations en utilisant les statistiques de Kruskal simplifiées $K'_{ji}, i \in 1, \ldots, C^R_{12}, j \in 1, \ldots, 142$. Les individus à classer sont ici les stations j et les variables sont les K_{ji}. Nous avons choisi un algorithme de classification très simple : Kmeans décrit dans [2]. Il a été testé pour un découpage en 4 saisons et pour une classification en 2, 3 et 4 classes. Les résultats sont présentés figure 4.12. On peut voir que pour un découpage en deux classes, une des classes est presque vide. On est donc tentés de dire qu'il n'y pas différents régimes de pluie dans la région. Lorsqu'on augmente le nombre de classes, deux régimes de pluie semblent se distinguer, un régime au nord et un régime au sud. Le découpage en saisons trouvé en moyenne pour chacune des classes est présenté dans les figures. On peut alors voir que les découpages trouvés pour le nord et le sud sont très proches. En conclusion, il existe une cohérence spatiale dans les classes obtenues par Kmeans, mais il ne semble pas y avoir différents régimes de pluie dans la région.

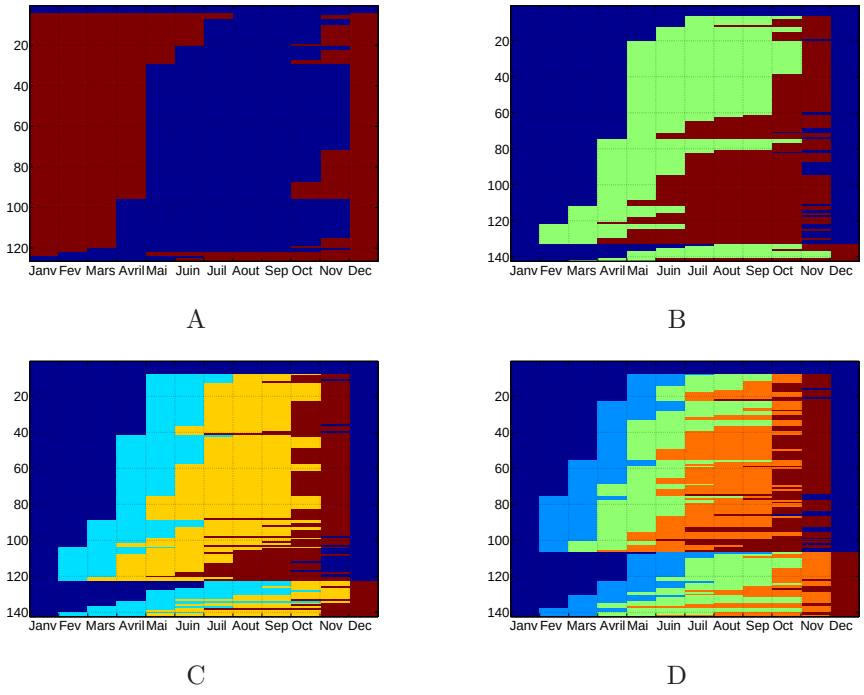

FIG. 4.11 – Saisons trouvées par l'approche Kruskal sur l'ensemble des stations. Le nombre de saisons a été fixé ici à 2, 3, 4 ou 5. Abscisses : mois. Ordonnées : Numéro de la station. Les mois d'une même saison sont représentés par une même couleur.

Idée de modèle

Avec l'approche Kruskal, il est difficile de déterminer avec certitude le découpage en saisons optimal. Nous avons donc chercher à développer un modèle plus souple

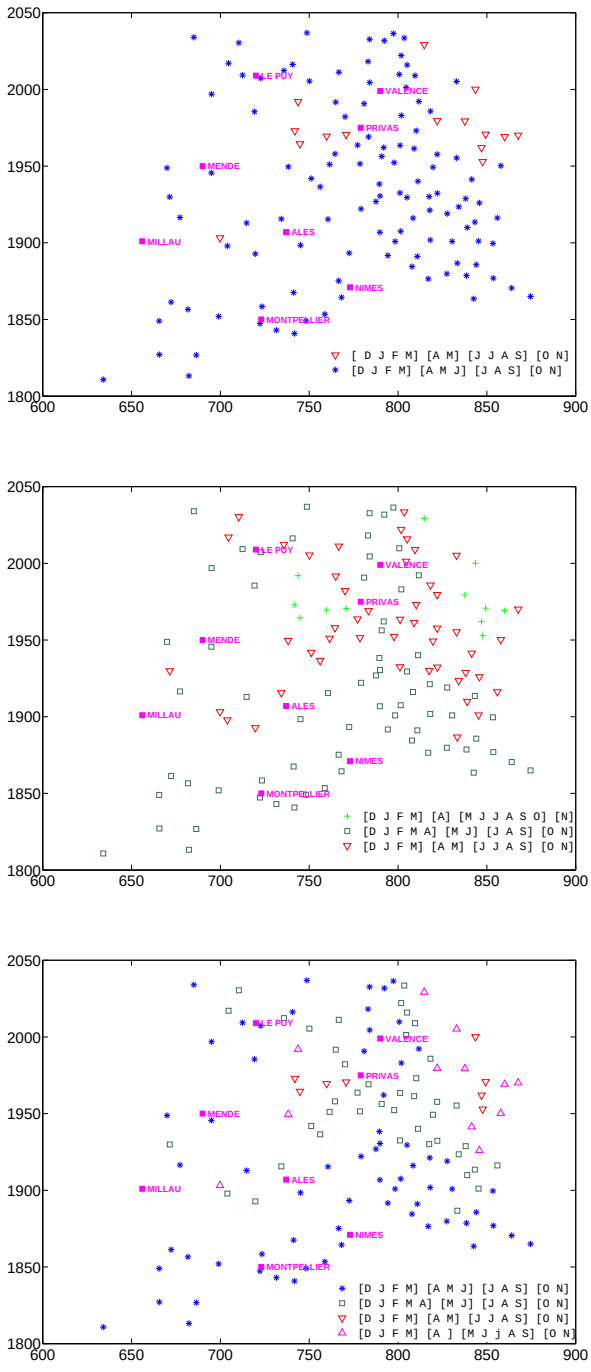

FIG. 4.12 – Classification des stations par Kmeans pour deux, trois et quatre classes (de haut en bas). Découpage en 4 saisons.

qu'un modèle de mélange de saisons. Nous présentons dans ce paragraphe l'idée du modèle sans validation. Soit X_t la hauteur de pluie au temps $t \in [0, A]$ (on oublie l'année). A représente une année. Il vaut 365 si on travaille à l'échelle de la journée, 365.25×24 si on travaille à l'échelle de l'heure.
Soit $u(t)$ le seuil en fonction du temps (déterministe).
Soit $D = \{(T, X_T), X_T > u(T)\}$ l'ensemble des dépassements X_T et des temps d'arrivée T de ces dépassements.
Soit $E = \{(T, Y_T = X_T - u(T)), X_T \in D\}$ l'ensemble des excès et de leurs temps d'arrivée.

Idée de la méthode L'idée du modèle proposé est la suivante : on considère que les dépassements suivent une loi GPD dont les paramètres de forme $\gamma(t)$ et d'échelle $\sigma(t)$ dépendent du temps. On considère de plus que ces dépassements arrivent avec une densité f. On décompose les paramètres $\gamma(t)$ et $\sigma(t)$ ainsi que la densité $f(t)$ sur une base de fonctions constantes par morceaux. Les paramètres introduits dans la décomposition sont alors estimés par maximisation de la vraisemblance.

Hypothèses du modèle

Hypothèse 1 : Modèle pour les excès On suppose que $Y(T)|T = t$ suit une loi GPD $(\gamma(t), \sigma(t))$, c'est à dire que :

$$\begin{aligned}\mathbb{P}(X_T - u(T) > x | T = t) &= \bar{G}(x, \gamma(t), \sigma(t)) \\ &= \mathbb{P}(X_T > x + u(T) | T = t)\end{aligned}$$

où $\bar{G}$ est la fonction de survie d'une loi GPD.
D'où :

$$\mathbb{P}(X_T > x | T = t) = \bar{G}(x - u(T), \gamma(t), \sigma(t)).$$

Hypothèse 2 : Modèle pour les temps de pluie on suppose que T est de densité f sur 365.25×24 heures.

On en déduit

$$\mathbb{P}(X_T > x) = \int_0^A \bar{G}(x - u(t), \gamma(t), \sigma(t)) f(t) dt.$$

Cette quantité correspond à la probabilité qu'un dépassement soit supérieur à x et non comme il a été vu precédement à la probabilité qu'une hauteur de pluie à un instant tiré aléatoirement soit supérieure à x. Il serait intéressant de voir avec les hydrologues quel est le calcul le plus intéressant pour eux.

Hypothèse 3 : Indépendance temps-excès On suppose que les temps d'arrivée des dépassements T et les dépassements X_T sont indépendants.

Estimation des paramètres par maximisation de la vraisemblance
Soit $\{(T_1, X_1), \ldots, (T_n, X_n)\}$ les dépassements observés. La vraisemblance s'écrit alors :

$$L = \prod_{i=1}^{n} g(X_i - u(T_i), \gamma(T_i), \sigma(T_i)) f(T_i).$$

D'où la log-vraisemblance :

$$\log L = \sum_{i=1}^{n} \log g(X_i - u(T_i, \gamma(T_i), \sigma(T_i)) + n \sum_{i=1} \log f(T_i)$$

On cherche alors à maximiser la log-vraisemblance ou de manière équivalente à minimiser la log-vraisemblance négative $-\log L$ avec la contrainte suivante sur f :

$$\int_0^A f(t)dt = 1.$$

Pour cela on introduit le Lagrangien λ :

$$\begin{aligned} -\log L &= -\sum_{i=1}^{n} \log g(X_i - u(T_i, \gamma(T_i), \sigma(T_i)) - \sum_{i=1}^{n} \log f(T_i) + \lambda \left(\int_0^A f(t)dt - 1 \right) \\ &= E + F \end{aligned}$$

avec

$$E = -\sum_{i=1}^{n} \log g(X_i - u(T_i, \gamma(T_i), \sigma(T_i))$$

et

$$F = -\sum_{i=1}^{n} \log f(T_i) + \lambda \left(\int_0^A f(t)dt - 1 \right).$$

Cela revient à résoudre deux problèmes de minimisations indépendants : un problème de minimisation par rapport aux fonctions γ et σ et un problème de minimisation en f.
L'idée de la méthode est de décomposer ces trois fonctions sur une base de fonctions $(e_k)_{k \geqslant 1}$. On se fixe un nombre $K \leqslant 12$ de fonctions.
On peut donc écrire :

$$f(t) = \sum_{k=1}^{K} f_k e_k(t).$$

Nous proposons de choisir les fonctions $(e_k)_{k \geqslant 1}$ constantes par mois mais on peut très bien décomposer f sur une base plus compliquée, par exemple une base de splines. Les fonctions de base s'écrivent donc :

$$e_k(t) = \sum_{l=1}^{12} a_{lk} \mathbb{1}_{t \in M_l}$$

D'où :

$$\begin{aligned} f(t) &= \sum_{k=1}^{K}\sum_{l=1}^{12} f_k a_{lk} \mathbb{1}_{t\in M_l} \\ &= \sum_{l=1}^{12}\sum_{k=1}^{K} a_{lk} f_k \mathbb{1}_{t\in M_l} \\ &= \sum_{l=1}^{12} b_l \mathbb{1}_{t\in M_l} \end{aligned}$$

où M_l désigne le mois l et $b_l = \sum_{k=1}^{K} a_{lk} f_k$

On a de plus la contrainte suivante pour tout $l = 1, \ldots, 12$:

$$\sum_{k=1}^{K} a_{lk} = 1.$$

Minimisation par rapport à f_k

$$\begin{aligned} F &= -\sum_{i=1}^{n} \log f(T_i) + \lambda \left(\int_0^A f(t)dt - 1 \right) \\ &= -\sum_{i=1}^{n} \log(\sum_{l=1}^{12} b_l \mathbb{1}_{T_i\in M_l}) + \lambda \left(\sum_{l=1}^{12} b_l \int_0^A \mathbb{1}_{u\in M_l} du - 1 \right) \\ &= -\sum_{i=1}^{n}\sum_{l=1}^{12} \log(b_l) \mathbb{1}_{T_i\in M_l} + \lambda \left(\sum_{l=1}^{12} b_l |M_l| - 1 \right) \\ &= -\sum_{l=1}^{12} n_l \log b_l + \lambda \left(\sum_{l=1}^{12} b_l |M_l| - 1 \right) \\ &= \sum_{l=1}^{12} (-n_l \log b_l + \lambda(b_l |M_l| - 1)) \end{aligned}$$

où n_l est le nombre d'excès qui tombent dans le mois M_l et $|M_l|$ est la taille du mois dans l'échelle considérée (heure, jour...). Pour minimiser B, on dérive par rapport à chaque f_j

$$\frac{\partial l(B)}{\partial f_j} = \sum_{l=1}^{12} (-n_l \frac{\partial \log b_l}{\partial f_j} + \lambda(\frac{\partial b_l}{\partial f_j}|M_l|)) \tag{4.3}$$

$$= \sum_{l=1}^{12} [-n_l \frac{a_{lj}}{b_l} + \lambda a_{lj}|M_l|]. \tag{4.4}$$

On multiplie l'équation (4.3) par f_j et on somme sur j. La minimisation de B par rapport à f_j conduit alors à la résolution des équations suivantes :

$$\sum_{l=1}^{12} n_l \frac{\sum_{j=1}^{K} f_j a_{lj}}{b_l} - \lambda \sum_{l=1}^{12}\sum_{j=1}^{K} f_j a_{lj} |M_l| = 0.$$

Comme $\sum_{l=1}^{12} n_l = n$ et $\sum_{l=1}^{12}\sum_{j=1}^{K} f_j a_{lj}|M_l| = 1$, on trouve que $\lambda = n$. Minimiser B par rapport à f_j revient donc à résoudre :

$$\sum_{l=1}^{12}[\frac{n_l a_{lj}}{\sum_{k=1}^{K} a_{lk} f_k} - n a_{lj}|M_l|] = 0, \forall j = 1, \ldots, K.$$

Notons que si $K = 12$ et $a_{lk} = \delta_{lk}$, on trouve $f_j = \frac{n_j}{n|M_j|}$.

Minimisation par rapport à γ et σ De la même manière que précédemment, on décompose $\gamma(t)$ et $\sigma(t)$ sur une base de fonctions constantes par morceaux, c'est à dire :

$$\gamma(t) = \sum_{k=1}^{K} \gamma_k e_k(t)$$

et

$$\sigma(t) = \sum_{k=1}^{K} \sigma_k e_k(t)$$

et

$$e_k(t) = \sum_{l=1}^{12} a_{lk} \mathbb{1}_{t \in M_l}.$$

On pose :

$$c_l = \sum_{k=1}^{K} a_{lk} \gamma_k$$

et

$$d_l = \sum_{k=1}^{K} a_{lk} \sigma_k.$$

On a alors :

$$\begin{aligned}
E &= -\sum_{i=1}^{n} \log(X_i - u(T_i, \gamma(T_i), \sigma(T_i)) \\
&= -\sum_{i=1}^{n} \log g(X_i - u(T_i), \sum_{l=1}^{12} c_l \mathbb{1}_{T_i \in M_l}, \sum_{l=1}^{12} d_l \mathbb{1}_{T_i \in M_l}) \\
&= -\sum_{i=1}^{n}\sum_{l=1}^{12} \log g(X_i - u(T_i), c_l, d_l) \mathbb{1}_{T_i \in M_l} \\
&= -\sum_{l=1}^{12} \sum_{T_i \in M_l} \log g(X_i - u(T_i), c_l, d_l).
\end{aligned}$$

On cherche donc à minimiser cette quantité A par rapport à γ_j et σ_j. Il n'y a pas de solution explicite.

Probabilité de dépassement de seuil, temps de retour Au final, la probabilité de dépasser un seuil x vaut :

$$\begin{aligned}\mathbb{P}(X_t > x) &= \int_0^A \bar{G}(x - u(t), \gamma(t), \sigma(t)) \sum_{l=1}^{12} \mathbb{1}_{t \in M_l} dt \\ &= \sum_{l=1}^{12} b_l \int_{M_l} \bar{G}(x - u(t), \gamma(t), \sigma(t)) dt\end{aligned}$$

On suppose que le seuil est constant et vaut u_l sur chaque mois M_l. On a alors :

$$\begin{aligned}\mathbb{P}(X_t > x) &= \sum_{l=1}^{12} b_l |M_l| \bar{G}(x - u_l, c_l, d_l) \\ &= \sum_{l=1}^{12} \sum_{k=1}^{K} a_{lk} f_k |M_l| \bar{G}(x - u_l, c_l, d_l) \\ &= \sum_{k=1}^{K} f_k [\sum_{l=1}^{12} a_{lk} |M_l| \bar{G}(x - u_l, c_l, d_l)].\end{aligned}$$

Remarque :
Notons que si on choisit $K = 12$ et $a_{lk} = \delta_{lk}$, alors on trouve que

$$\mathbb{P}(X_t > x) = \sum_{k=1}^{12} \frac{n_k}{n} \bar{G}(x - u_k, \gamma_k, \sigma_k)$$

avec $n = \sum_{l=1}^{12} n_l$ On ne retrouve pas la probabilité $\mathbb{P}(X_t > x)$ du modèle de mélange (voir équation (4.2)) car ici n est le nombre de dépassements sur l'année et non le nombre de mesures sur l'année. Cependant, il est normal de ne pas retrouver la même chose puisque dans un cas, on calcule la probabilité qu'un dépassement soit supérieur à x et dans l'autre cas, on calcule la probabilité qu'une hauteur d'eau quelconque soit supérieure à x. Il serait peut être intéressant de faire le calcul de cette dernière quantité dans le cadre de ce modèle pour voir si les résultats sont identiques à un modèle de mélange.

Chapitre 5

Conclusion

Nous avons présenté dans ce rapport une première analyse des pluies extrêmes dans la région des Cévennes-Vivarais. Les principales conclusions de cette analyse sont les suivantes :

- Le domaine de Fréchet semble le plus apte à modéliser les queues de distribution des hauteurs de pluie.
- Les paramètres de la loi GPD utilisée pour modéliser les queues de distribution dépendent fortement du relief. De même pour les temps et niveaux de retour.
- En fonction du pas de temps choisi pour les mesures, les cartes des temps et niveaux de retour sont radicalement différentes. Pour un cumul de pluie horaire, les pluies intenses se situent davantage en plaine alors que pour un cumul journalier, elles se situent sur les sommets.
- Une forte saisonnalité des mesures a été mise en évidence et devrait être prise en compte dans les estimations des temps et niveaux de retour. Nous avons d'ailleurs proposé pour ce faire un modèle basé sur le découpage des séries chronologiques en saisons. Le choix des saisons peut être réalisé par une approche non paramétrique basée sur la statistique de Kruskal-Wallis.
- Enfin, il semble qu'une corrélation temporelle existe dans les séries chronologiques.

Pour la suite, nous pensons qu'il serait intéressant de tester dans un premier temps le modèle que nous avons proposé. La question de savoir si il faut prendre en compte la corrélation temporelle dans ce modèle reste ouverte. Enfin, le développement d'un modèle spatial reste l'objectif final de cette étude. A ce sujet, nous renvoyons le lecteur aux différents papiers écrits récemment sur le sujet [28, 25, 11, 15, 1, 16, 12, 4].

Annexe A

Programmes

Pour télécharger les données de pluie et ajouter l'ensemble des programmes et packages au path de matlab, il faut lancer le programme loadHYDRO.m qui se trouve dans le répertoire DOSSIERHYDRO/ :
>> loadHYDRO

L'ensemble des programmes permettant de retrouver tous les résultats de ce rapport sont disponibles dans le répertoire :
DOSSIERHYDRO/ProgHYDRO/Chap1/
Les programmes utilisent deux packages Matlab :

- Le package EVIM pour l'étude des valeurs extrêmes [17]
- Le package de géostatistique BMElib [7]

LISTE DES PROGRAMMMES :

ajoutvilles.m : Place les villes de Valence, Privas, Ales, Nîmes, Montpellier, Millau, Mende et Le Puy sur une carte. Les cinq principaux sommets de la région Cévennes-Vivarais sont indiqués par un triangle sur la carte.
>> *ajoutvilles*

AnalSpatialHill.m : Produit le variogramme, la carte d'estimation par krigeage et la carte de variance d'estimation pour :

- γ le paramètre de forme de la loi GPD estimé par l'estimateur de Hill.
- σ le paramètre d'échelle de la loi GPD déduit en multipliant γ par le seuil
- le temps de retour pour une intensité donnée
- le niveau de retour pour un temps de retour donné

AnalSpatialHill(stations,dates,mesures,pourcentage,temps de retour,intensité) ;

- stations : coordonnées des stations. Matrice de taille $(n \times 2)$
- dates : dates des mesures. Vecteur de taille n.
- mesures : mesures pour les p stations. Matrice de taille $(n \times p)$
- pourcentage : pourcentage d'excès à conserver pour l'étude
- temps de retour : temps de retour pour lequel on souhaite estimer les niveaux de retour
- intensité : intensité pour laquelle on souhaite estimer les temps de retour

Exemple :
>> res=AnalSpatialHill(stationsXYZ,datesSelect,mesuresSelect,10,10,50) ; Matlab affiche alors le variogramme expérimental pour le paramètre γ. Par défaut un modèle de type sphérique avec effet de pépite est ajusté. Il faut cependant que l'utilisateur initialise les paramètres, d'où le message suivant de Matlab :

Please give the range and sill for each model :

L'utilisateur doit alors donner une première valeur pour l'effet de pépite, la portée et le palier du modèle sphérique sous la forme suivante :
{*[pépite],[portée sphérique palier sphérique]*}
par exemple :
{*[0.001],[0.003 150]*}
S'affiche alors le variogramme expérimental et le modèle ajusté, la cartographie des γ par krigeage et la variance d'estimation associée. Le programme continue dans la même logique avec le paramètre d'échelle σ, les temps de retour et les niveaux de retour.

L'ensemble des résultats est stocké sous forme de liste :
res =

- nbannees : nombre d'années de mesures pour chaque station
- seuil : seuil retenu pour chaque station
- nbseuil : nombre de mesures retenues pour chaque station
- gamma : estimations de γ pour chaque station
- sigma : estimations de σ pour chaque station
- intens : intensité estimée pour chaque station pour le temps de retour donné en entrée de programme
- TmpsEstim : période de retour pour chaque station pour l'intensité donnée en entrée de programme
- resVkrigGAMMA : carte de la variance de krigeage des γ
- reskrigGAMMA : carte des γ estimés par krigeage
- resVkrigSIGMA : carte de la variance de krigeage des σ
- reskrigSIGMA : carte des σ estimés par krigeage
- resVkrigTMPS : carte de la variance de krigeage des temps de retour
- reskrigTMPS : carte des temps de retour estimés par krigeage
- resVkrigINTENS : carte de la variance de krigeage des niveaux de retour
- reskrigINTENS : carte des niveaux de retour estimés par krigeage
- x : abscisse pour chaque station
- y : ordonnée pour chaque station
- paramfitGAMMA : paramètres retenus pour le modèle variographique des γ
- paramfitSIGMA : paramètres retenus pour le modèle variographique des σ
- paramfitTMPS : paramètres retenus pour le modèle variographique des temps de retour
- paramfitINTENS : paramètres retenus pour le modèle variographique des niveaux de retour

AnalSpatialML.m : Même programme que AnalSpatialHill pour des estimations par maximum de vraisemblance
Exemple :
>> res=AnalSpatialML(stationsXYZ,datesSelect,mesuresSelect,10,10,50) ;

ChoixSeuilFrechet.m :Calcule pour une station donnée, en fonction du pourcentage d'excès retenus, la p-valeur des tests d'adéquation des variables

$$Z_i = \log \frac{Y_i}{u}$$

à une loi exponentielle (Chi 2 et Anderson-Darling, voir section 3.1.1)

[pAnders, pChi2,thres,k]=ChoixSeuilFrechet(pourcentage,mesures)
- pourcentage :pourcentage d'excès à conserver pour l'étude
- mesures : mesures pour une station (vecteur de taille n)
- pAnders : p-valeur pour le test d'Anderson-Darling
- pChi2 : p-valeur pour le test du Chi2
- thres : seuil
- k : nombre de mesures au dessus du seuil

Exemple :
>> [pAnders, pChi2,thres,k]=ChoixSeuilFréchet(50,mesuresSelect(:,4))

ChoixSeuilGPD.m : Calcule pour une station donnée, en fonction du pourcentage d'excès retenus, la p-valeur du test du Chi2 d'adéquation des excès à une loi GPD, voir section 3.1.1
S'utilise de la même manière que ChoixSeuilFréchet
Exemple :
>> [pChi2,thres,k]=ChoixSeuilGPD(50,mesuresSelect(:,4))

ConvertDataMat.m : Groupe les mesures par mois pour une station donnée. Le résultat de cette fonction est une matrice de taille ($n \times 12$)
ConvertDataMat(station,datesSelect,mesuresSelect)
- station : station d'étude
- dates : dates de mesure, vecteur de dimension n
- mesures : mesures, vecteur de dimension n

Exemple :
>> ConvertDataMat(station,datesSelect,mesuresSelect)

CumulDay.m : Calcule la hauteur de pluie cumulée par jour à partir des données horaires

[resSize,resCumul,resDate]=CumulDay(dates, mesures)
- mesures : mesures horaires pour une station (vecteur de taille n)
- dates : dates des mesures
- resSize : nombre de mesures positives pour chaque jour
- resCumul : hauteur de pluie cumulée par jour
- resDate : dates

Exemple :
>> [resSize,resMean,resDate]=CumulDay(datesSelect, mesuresSelect(:,4))

CumulMonth.m : Calcule la hauteur de pluie cumulée par mois à partir des données horaires

[resSize,resCumul]=CumulMonth(dates, mesures)

- mesures : mesures horaires pour une station (vecteur de taille n)
- dates : dates des mesures
- resSize : nombre de mesures positives chaque mois
- resCumul : hauteur de pluie cumulée par mois

Exemple :
>> [resSize,resCumul]=CumulMonth(dates, measures)

CumulYear.m : Calcule la hauteur de pluie cumulée par an à partir de données horaires

[resSize,resCumul]=CumulYear(dates, mesures)

- mesures : mesures horaires pour une station (vecteur de taille n)
- dates : dates des mesures
- resSize : nombre de mesures positives par année
- resCumul : hauteur de pluie cumulée par année

Exemple :
>> [resSize,resMean]=CumulYear(datesSelect, mesuresSelect(:,4))

EstimHill.m : Estimations du paramètre de forme par Hill avec intervalles de confiance à 95%

[gamma,ci]=EstimHill(pourcentage,mesures)

- pourcentage :pourcentage d'excès à conserver pour l'étude
- mesures : mesures pour une station (vecteur de taille n)
- gamma : estimation du paramètre de forme
- ci : bornes min et max de l'intervalle de confiance à 95% associé

Exemple :
>> [res,ci]=EstimHill(50,mesuresSelect(:,4))

EstimML.m : Estimations des paramètres de forme et d'échelle par maximum de vraisemblance avec intervalles de confiance à 95%

[gamma,sigma]=EstimML(pourcentage,mesures)

- pourcentage :pourcentage d'excès à conserver pour l'étude
- mesures : mesures pour une station (vecteur de taille n)
- gamma : estimation du paramètre de forme suivi des bornes min et max de l'intervalle de confiance à 95% associé
- sigma : estimation du paramètre d'échelle suivi des bornes min et max de l'intervalle de confiance à 95% associé

Exemple :
>> [gamma,sigma]=*EstimML(50,mesuresSelect(:,4))*

ExtraitPercentSaison.m : Extrait les dépassements (le pourcentage est fixé par l'utilisateur) pour une saison donnée.
ExtraitPercentSaison(Mesures,mois,percent)

- Mesures : les mesures doivent être groupées par mois. On les entre sous la forme d'une matrice de taille $(n \times 12)$ dans laquelle chaque colonne j contient les n mesures du mois j. Il arrive fréquemment que le nombre de mesures diffère par mois. Dans ce cas les mesures manquantes seront remplacées par des NaN. Nous proposons d'utiliser la fonction *Convert-DataMat* pour convertir les mesures initiales dans le format demandé.

- mois : numéros des mois de la saison
- percent : pourcentage de dépassements à conserver pour l'étude

Exemple :
>> matessai=ConvertDataMat(station,datesSelect,mesuresSelect) ;
>> ExtraitPercentSaison(matessai,[1 2 3],10)

Hillparmois.m : Estimation du paramètre de forme par Hill par mois

[gamma,ci]=Hillparmois(pourcentage,dates,mesures)
- pourcentage :pourcentage d'excès à conserver pour l'étude
- dates : dates des mesures (vecteur de taille n)
- mesures : mesures pour une station (vecteur de taille n)
- gamma : estimation du paramètre de forme par mois
- ci : bornes de l'intervalle de confiance à 95% par mois

Exemple :
>> [gamma,ci]=Hillparmois(50,datesSelect,mesuresSelect(:,4))

hillplot2.m : Estimation du paramètre de forme par Hill en fonction du pourcentage d'excès retenus pour l'étude et intervalle de confiance à 95%

[gamma,uband,lband]=hillplot2(mesures,'xi','n',t) ;
- mesures : mesures pour une station (vecteur de taille n)
- gamma : estimation du paramètre de forme en fonction du pourcentage d'excès
- t : niveau de confiance pour l'intervalle de confiance
- uband : borne supérieure de l'intervalle de confiance à t%
- lband : borne inférieure de l'intervalle de confiance à t%

Exemple :
>> [c,uband,lband]=hillplot2(mesuresSelect(:,4),'xi','n',0.95) ;

MeanDay.m : Calcule l'intensité (hauteur de pluie /heure) de pluie moyenne par jour

[resSize,resMean,resDate]=MeanDay(dates, mesures)
- mesures : mesures horaires pour une station (vecteur de taille n)
- dates : dates des mesures (vecteur de taille n)
- resSize : nombre de mesures positives pour chaque jour
- resMean : intensité moyenne de pluie par jour
- resDate : dates (dates des jours à minuit, utile pour tracer la chronique par exemple)

Exemple :
>> [resSize,resMean,resDate]=MeanDay(datesSelect, mesuresSelect(:,4))

MeanMonth.m : Calcule l'intensité de pluie moyenne par mois

[resSize,resMean]=MeanMonth(dates, mesures)
- mesures : mesures horaires pour une station (vecteur de taille n)
- dates : dates des mesures (vecteur de taille n)
- resSize : nombre de mesures positives chaque mois
- resMean : intensité moyenne de pluie par mois

Exemple :
>> [resSize,resMean]=MeanMonth(dates, measures)

MeanYear.m : Calcule l'intensité de pluie moyenne par an

[resSize,resMean]=MeanYear(dates, mesures)
- mesures : mesures horaires pour une station (vecteur de taille n)
- dates : dates des mesures (vecteur de taille n)
- resSize : nombre de mesures positives par année
- resMean : intensité moyenne de pluie par an

Exemple :
>> [resSize,resMean]=MeanYear(datesSelect, mesuresSelect(:,4))

maxDay.m : Calcule la hauteur de pluie horaire maximale par jour

[resSize,resMax,resDate]=maxDay(datesSelect, mesuresSelect(:,4)) ;
- mesures : mesures horaires pour une station (vecteur de taille n)
- dates : dates des mesures (vecteur de taille n)
- resSize : nombre de mesures positives pour chaque jour
- resMax : hauteur de pluie (horaire) maximale par jour
- resDate : dates (date-heure à laquelle le maximum est observé)

Exemple :
>> [resSize,resMax,resDate]=MaxDay(datesSelect, mesuresSelect(:,4))

maxMonth.m : Calcule la hauteur de pluie horaire maximale par mois

[resSize,resMax,resDate]=maxMonth(dates, mesures) ;
- mesures : mesures horaires pour une station (vecteur de taille n)
- dates : dates des mesures (vecteur de taille n)
- resSize : nombre de mesures positives pour chaque mois
- resMax : hauteur de pluie (horaire) maximale par mois
- resDate : dates (date-heure à laquelle le maximum est observé)

Exemple :
>> [resSize,resMax,resDate]=MaxDay(datesSelect, mesuresSelect(:,4))

maxYear.m : Calcule la hauteur de pluie horaire maximale par année

[resSize,resMax,resDate]=maxYear(dates, mesures) ;
- mesures : mesures horaires pour une station (vecteur de taille n)
- dates : dates des mesures (vecteur de taille n)
- resSize : nombre de mesures positives pour chaque année
- resMax : hauteur de pluie (horaire) maximale par année
- resDate : dates (date-heure à laquelle le maximum est observé)

Exemple :
>> [resSize,resMax,resDate]=MaxYear(datesSelect, mesuresSelect(:,4))

MLparmois.m : Estimations des paramètres de forme et d'échelle par maximum de vraisemblance par mois

[gamma,sigma]=MLparmois(pourcentage,dates,mesures) avec
- pourcentage :pourcentage d'excès à conserver pour l'étude
- dates : dates des mesures (vecteur de taille n)
- mesures : mesures pour une station (vecteur de taille n)
- gamma : estimation du paramètre de forme et intervalle de confiance à 95% (par mois). La $i$ème ligne correspond au $i$ème mois.
- sigma : estimation du paramètre d'échelle et intervalle de confiance à 95% (par mois). La $i$ème ligne correspond au $i$ème mois.

Exemple :
>> [gamma,sigma]=MLparmois(50,datesSelect,mesuresSelect(:,4))

MLplot.m : Estimation des paramètres de forme et d'échelle en fonction du pourcentage d'excès retenus pour l'étude et intervalles de confiance à 95%

[gamma,sigma]=MLplot(mesures,pourcentage)
- mesures : mesures pour une station (vecteur de taille n)
- pourcentage :pourcentage d'excès à conserver pour l'étude
- gamma : estimation du paramètre de forme et intervalle de confiance à 95% (par mois). La $i$ème ligne correspond à i % d'excès.
- sigma : estimation du paramètre d'échelle et intervalle de confiance à 95% (par mois). La $i$ème ligne correspond à i % d'excès.

Exemple :
>> [gamma,sigma]=MLplot(mesuresSelect(:,4),50) ;

SeasonsCostsKruskalPmois.m : Calcule la statistique de Kruskal-Wallis simplifiée pour une configuration de saison donnée. Avec cette fonction on se fixe un pourcentage d'excès par mois et non par saison.
SeasonsCostsKruskalPmois(Config,mesures,percent)
- Config : vecteur de taille R décrivant les R saisons. Si config vaut par exemple [2 5 8 10], c'est qu'il y a quatre saisons. La première saison est constituée des mois de février, mars et avril, la deuxième de mai, juin et juillet, la troisième d'août, septembre et la quatrième d'octobre, novembre, décembre et janvier. Chaque composante du vecteur indique le premier mois de la nouvelle saison. Notons que cette écriture est compatible avec la fonction nchoosek de matlab qui décrit les C_n^k combinaisons possibles de k valeurs parmi n. Cette fonction nous est utile pour décrire toutes les combinaisons de saisons possibles.
- Mesures : les mesures doivent être groupées par mois. On les entre sous la forme d'une matrice de taille $(n \times 12)$ dans laquelle chaque colonne j contient les n mesures du mois j. Il arrive fréquemment que le nombre de mesures diffère par mois. Dans ce cas les mesures manquantes seront remplacées par des NaN. Nous proposons d'utiliser la fonction *Convert-DataMat* pour convertir les mesures initiales dans le format demandé.
- Percent : pourcentage d'excès retenus par mois

Exemple :
Pour une station donnée, pour calculer la statistique de Kruskal-Wallis sur l'ensemble des configurations possibles pour R saisons, on procède ainsi :
>> station=99 ;
>> R=4 ;

>> percent=10 ;
>> config=nchoosek(1 :12,R) ;
>> for (k=1 :length(config))
>> matessai=ConvertDataMat(station,datesSelect,mesuresSelect) ;
>> res(k)=SeasonsCostsKruskalPmois(config(k, :),matessai,percent) ;
>> end
Pour trouver la saison la plus pertinente, il suffit alors de chercher pour quelle saison res est maximal :
>> ind=find(res==max(res)) ;
>> config(ind, :)
Dans cet exemple on trouve par exemple : [D J F M A] [M] [J J A S] [O N]

SeasonsCostsKruskalPsaison.m : Calcule la statistique de Kruskal-Wallis simplifiée pour une configuration de saison donnée. Avec cette fonction on se fixe un pourcentage par saison.
SeasonsCostsKruskalPmois(Config,mesures,percent)

- Config : vecteur de taille R décrivant les R saisons. Si config vaut par exemple [2 5 8 10], c'est qu'il y a quatre saisons. La première saison est constituée des mois de février, mars et avril, la deuxième de mai, juin et juillet, la troisième d'août, septembre et la quatrième d'octobre, novembre, décembre et janvier. Chaque composante du vecteur indique le premier mois de la nouvelle saison. Notons que cette écriture est compatible avec la fonction nchoosek de matlab qui décrit les C_n^k combinaisons possibles de k valeurs parmi n. Cette fonction nous est utile pour décrire toutes les combinaisons de saisons possibles.
- Mesures : les mesures doivent être groupées par mois. On les rentre sous la forme d'une matrice de taille $(n \times 12)$ dans laquelle chaque colonne j contient les n mesures du mois j. Il arrive fréquemment que le nombre de mesures diffère par mois. Dans ce cas les mesures manquantes seront remplacées par des NaN. Nous proposons d'utiliser la fonction *Convert-DataMat* pour convertir les mesures initiales dans le format demandé.
- Percent : pourcentage d'excès retenus par mois

Exemple :
Pour une station donnée, pour calculer la statistique de Kruskal-Wallis sur l'ensemble des configurations possibles pour R saisons, on procède ainsi :
>> station=99 ;
>> R=4 ;
>> percent=10 ;
>> config=nchoosek(1 :12,R) ;
>> for (k=1 :length(config))
>> matessai=ConvertDataMat(station,datesSelect,mesuresSelect) ;
>> res(k)=SeasonsCostsKruskalPmois(config(k, :),matessai,percent) ;
>> end
Pour trouver la saison la plus pertinente, il suffit alors de chercher pour quelle saison res est maximal :
>> ind=find(res==max(res)) ;
>> config(ind, :)
Dans cet exemple on trouve par exemple : [J F M A] [M J J A] [S O] [N D]

SeasonsCostsPVAL.m : Effectue le test de kruskal-Wallis dans chaque saison pour vérifier que les mesures par mois suivent bien la même loi. Renvoie (1 - le min des p-valeurs sur les R saisons).
SeasonsCostsPVAL(config,mesures,percent,theta)

- config : vecteur de taille R décrivant les R saisons. Si config vaut par exemple [2 5 8 10], c'est qu'il y a quatre saisons. La première saison est constituée des mois de février, mars et avril, la deuxième de mai, juin et juillet, la troisième d'août, septembre et la quatrième d'octobre, novembre, décembre et janvier. Chaque composante du vecteur indique le premier mois de la nouvelle saison. Notons que cette écriture est compatible avec la fonction nchoosek de matlab qui décrit les C_n^k combinaisons possibles de k valeurs parmi n. Cette fonction nous est utile pour décrire toutes les combinaisons de saisons possibles.
- mesures : les mesures doivent être groupées par mois. On les entre sous la forme d'une matrice de taille $(n \times 12)$ dans laquelle chaque colonne j contient les n mesures du mois j. Il arrive fréquemment que le nombre de mesures diffère par mois. Dans ce cas les mesures manquantes seront remplacées par des NaN. Nous proposons d'utiliser la fonction *ConvertDataMat* pour convertir les mesures initiales dans le format demandé.
- Percent : pourcentage d'excès retenus par mois
- theta : il faut fixer theta à 1

Exemple :
Pour une station donnée, pour calculer la statistique de Kruskal-Wallis sur l'ensemble des configurations possibles pour R saisons, on procède ainsi :

```
>> station=99 ;
>> R=4 ;
>> percent=10 ;
>> config=nchoosek(1 :12,R) ;
>> for (k=1 :length(config))
>> matessai=ConvertDataMat(station,datesSelect,mesuresSelect) ;
>> res(k)=SeasonsCostsPVAL(config(k, :),matessai,percent,1) ;
>> end
```

Pour trouver la saison la plus pertinente, il suffit alors de chercher pour quelle saison res est maximal :

```
>> ind=find(res==min(res)) ;
>> config(ind, :)
```

Dans cet exemple on trouve par exemple : [J F M A] [M J J A S] [O N] [D]

TestFrechet.m : Diagramme de Hill
TestFrechet(pourcentage,mesures)

- mesures : mesures pour une station (vecteur de taille n)
- pourcentage :pourcentage d'excès à conserver pour l'étude

Exemple :

```
>> TestFrechet(50,mesuresSelect( :,4))
```

TestGumbel.m : Return level plot
TestGumbel(mesures)

- mesures : mesures pour une station (vecteur de taille n)

Exemple :
>> TestGumbel(resMax(:,4))

VarioTemp.m : Calcule le variogramme expérimental temporel des excès pour une station
VarioTemp(dates,mesures,percent,(24*30),(12*5))

- dates (vecteur de taille n)
- mesures : mesures pour une station (vecteur de taille n)
- percent : pourcentage d'excès à retenir. Si on souhaite conserver toutes les mesures, percent doit être égal à 100%
- pas : pas de calcul du variogramme en heures
- taille : choix du nombre de pas

Exemple :
*>> VarioTemp(datesSelect,mesuresSelect(:,40),100,(24*30),(12*5))*
Trace le variogramme pour un pas de temps d'environ un mois

Annexe B

Carte du relief

FIG. B.1 – Carte du relief dans la région Cévennes-Vivarais

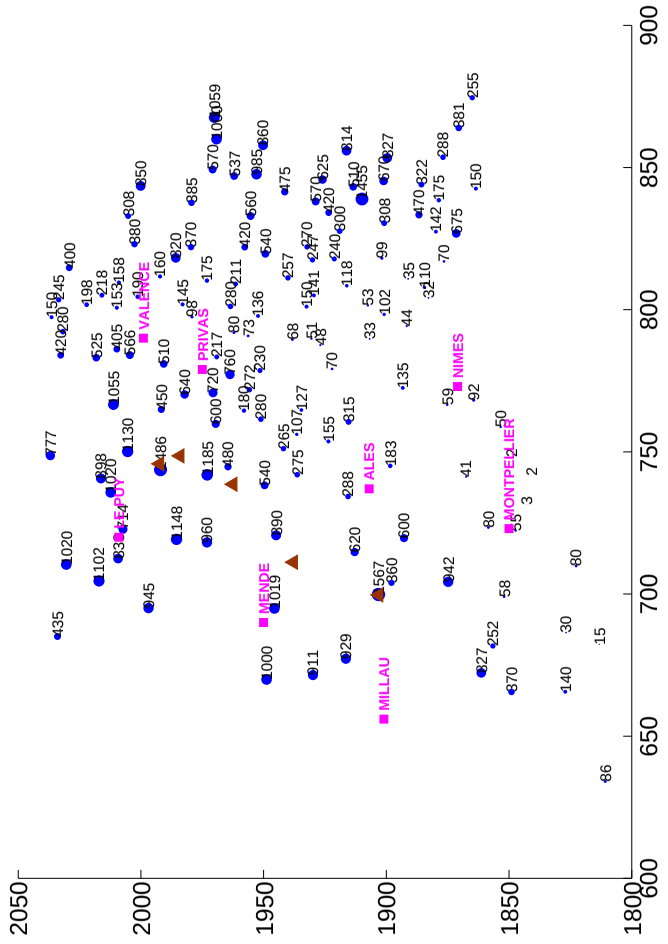

FIG. B.2 – Altitude des stations

Annexe C

Choix du seuil, données horaires

Trois critères ont été étudiés pour le choix du pourcentage d'excès :
- Tests d'adéquation des $Z_i = \log(Y_i/u)$ à la loi exponentielle,
- Evolution des paramètres de forme et d'échelle (par maximum de vraisemblance et par Hill) en fonction du pourcentage d'excès retenu,
- Cohérence des estimations par maximum de vraisemblance et par Hill pour les paramètres de forme et d'échelle.

Les résultats sont présentés dans cette annexe pour des données horaires et pour les trois stations les mieux informées du réseau : à Barnas, à Mazan l'Abbaye et à Valleraugue.

Les tests d'adéquation à la loi exponentielle (figures C.1) montrent pour ces trois stations qu'à partir de 3-4% d'excès, l'hypothèse d'une loi exponentielle pour les Z_i n'est plus valable. En effectuant cette étude pour l'ensemble des 126 stations et en retenant pour chacune le seuil minimal à partir duquel la p-valeur est inférieure à 5%, on trouve qu'en moyenne, le pourcentage d'excès doit être fixé à 4% (par Anderson-Darling et par Chi2). Ce pourcentage d'excès correspond à 84 excès en moyenne par station et à un seuil moyen de 9mm. Les résultats obtenus par le test du Chi 2 ou par le test d'Anderson-Darling sont très similaires.

L'évolution des paramètres de forme et d'échelle en fonction du pourcentage d'excès est difficile à analyser. Par maximum de vraisemblance (figure C.2), les estimations semblent se stabiliser à partir de 10-20% d'excès. Par l'estimateur de Hill (figure C.3), les estimations sont stables pour un pourcentage d'excès autour de 4-5%. Dans les deux cas, il est difficile de fixer un seuil.

Enfin, la comparaison des estimations du paramètre de forme par maximum de vraisemblance et par Hill (figure C.4) montre que pour les stations de Barnas et de Mazan L'Abbaye, avec plus 5-6 % d'excès, les estimations ne sont plus cohérentes entre les deux méthodes, ce qui signifie que le modèle n'est probablement plus valable. Pour la station de Valleurauge, on peut toutefois conserver jusqu'à 15% d'excès pour que les estimations restent cohérentes. Globalement, un pourcentage d'excès fixé à 4% semble approprié pour conserver une cohérence entre les estimations du paramètre de forme par maximum de vraisemblance et par Hill.

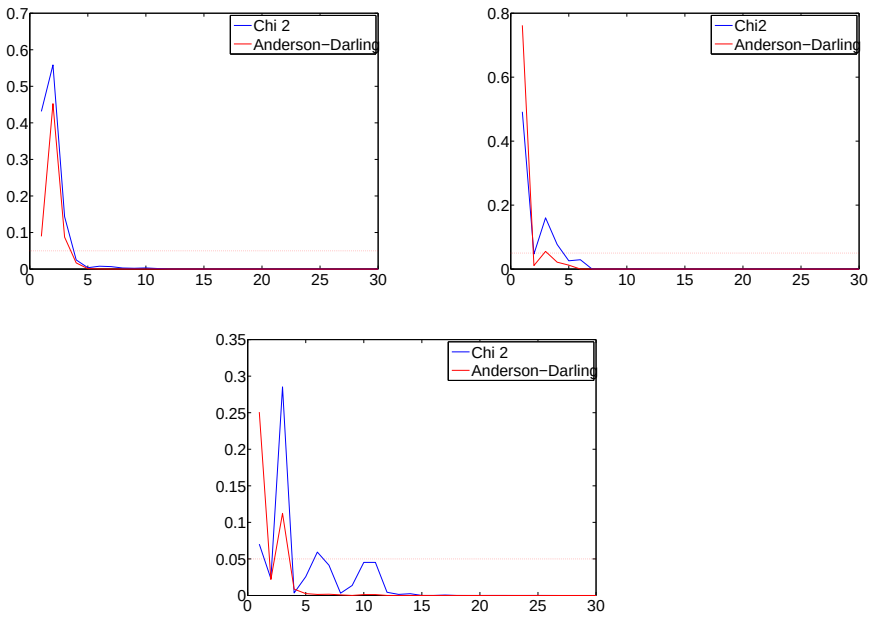

FIG. C.1 – Tests d'adéquation des Z_i à la loi exponentielle. p-valeur en fonction du pourcentage d'excès. Données horaires.

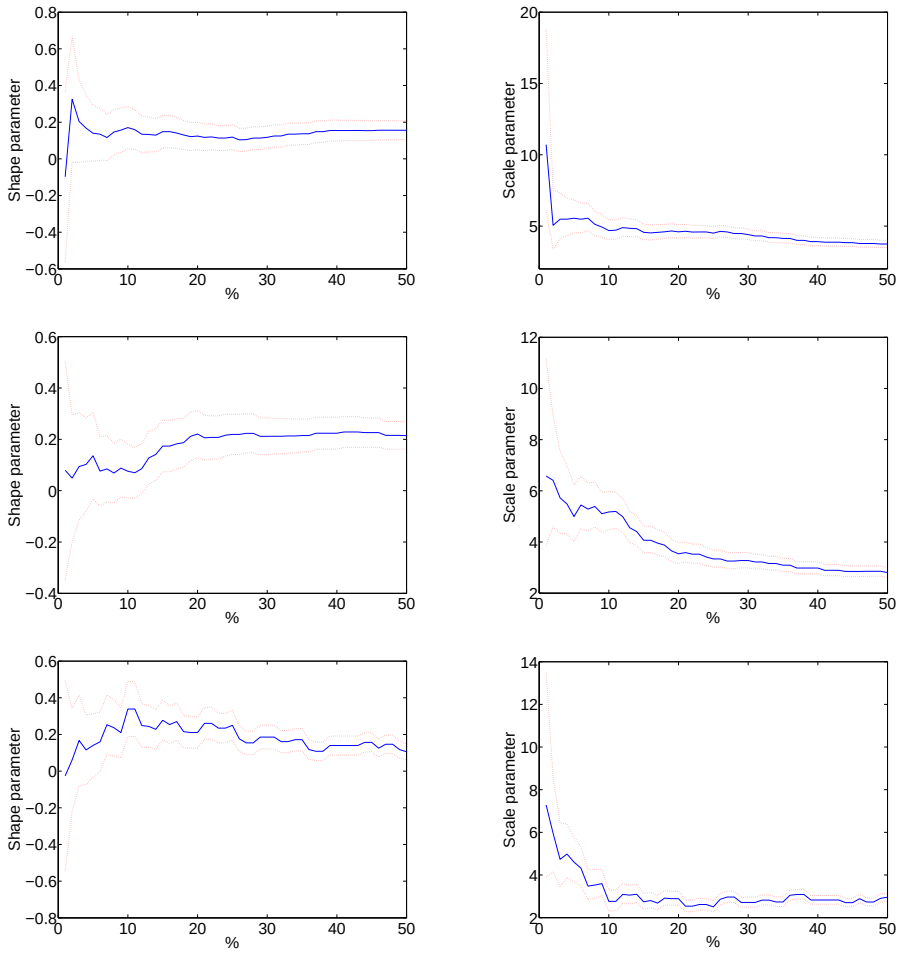

FIG. C.2 – Estimation des paramètre de forme (à gauche) et d'échelle (à droite) par maximum de vraisemblance et intervalles de confiance. En haut à gauche : à Barnas, En haut à droite : à Mazan-L'Abbaye. En bas : A Valleraugue. Données horaires.

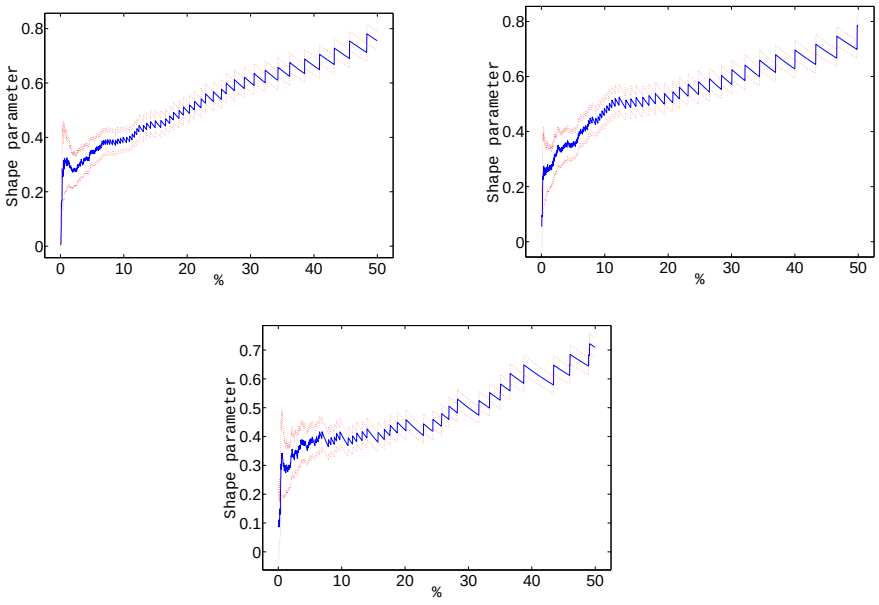

FIG. C.3 – Estimation des paramètre de forme par Hill et intervalles de confiance. En haut à gauche : à Barnas, En haut à droite : à Mazan-L'Abbaye. En bas : A Valleraugue. Données horaires.

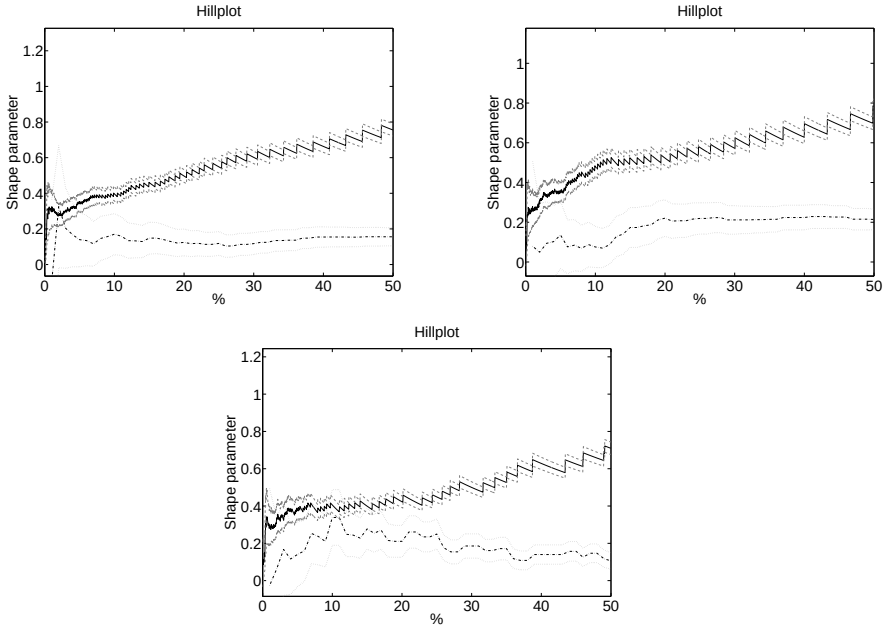

FIG. C.4 – Cohérence des estimations du paramètre de forme par maximum de vraisemblance (trait noir en pointillé) et par Hill (trait noir continu) pour les station de Barnas (en haut à gauche), de Mazan-L'Abbaye (en haut à droite) et de Valleraugue (en bas). Données horaires.

Annexe D

Choix du seuil, données journalières

Pour des données journalières, les tests d'adéquations (figures D.1) sur les trois stations les mieux informées montrent qu'un pourcentage d'excès compris entre 5 et 15% est à envisager. Sur l'ensemble des stations, le pourcentage d'excès est fixé à 9% en moyenne par le test du Chi2 ce qui correspond à un seuil moyen de 32 mm par jour et à en moyenne 47 mesures par station et à 4% par le test d'Anderson, ce qui correspond à un seuil moyen de 47 mm par jour et à en moyenne 20 mesures par station.
Comme pour les données horaires, il est difficile de choisir le seuil en regardant l'évolution des estimations des paramètres en fonction du pourcentage d'excès (figures D.2 et D.3). Pour la station de Valleraugue, on note toutefois une certaine stabilité des estimations pour l'approche Hill autour de 10% d'excès.
Enfin, la comparaison des estimations par Hill et par maximum de vraisemblance montrent qu'après 10% d'excès les estimations ne sont plus cohérentes (5% d'excès pour la station de Valleraugue).
Globalement, un pourcentage d'excès fixé à 10% nous semble approprié pour cette étude. Il correspond à un seuil moyen de 30 mm par jour et à en moyenne 52 mesures par station.

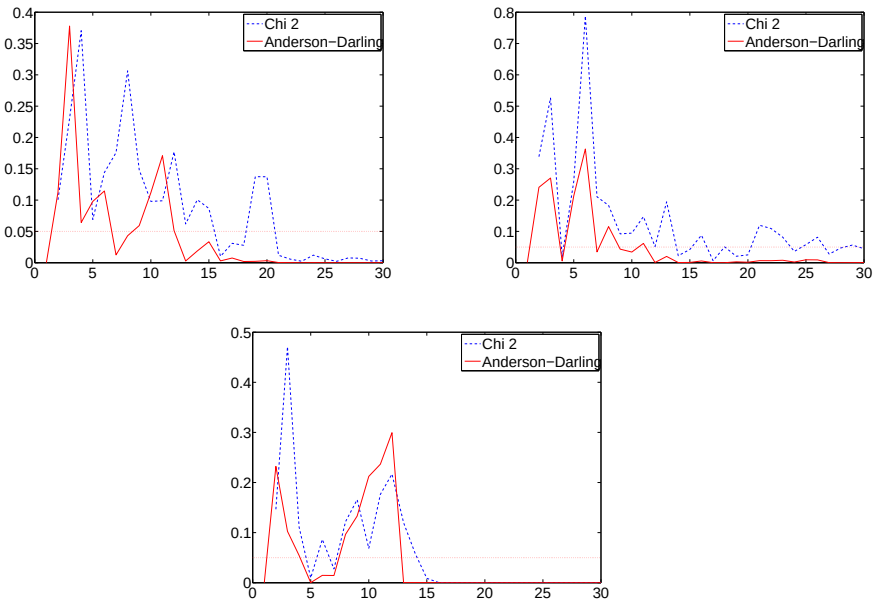

FIG. D.1 – Tests d'adéquation des Z_i à la loi exponentielle. p-valeur en fonction du pourcentage d'excès. Données journalières.

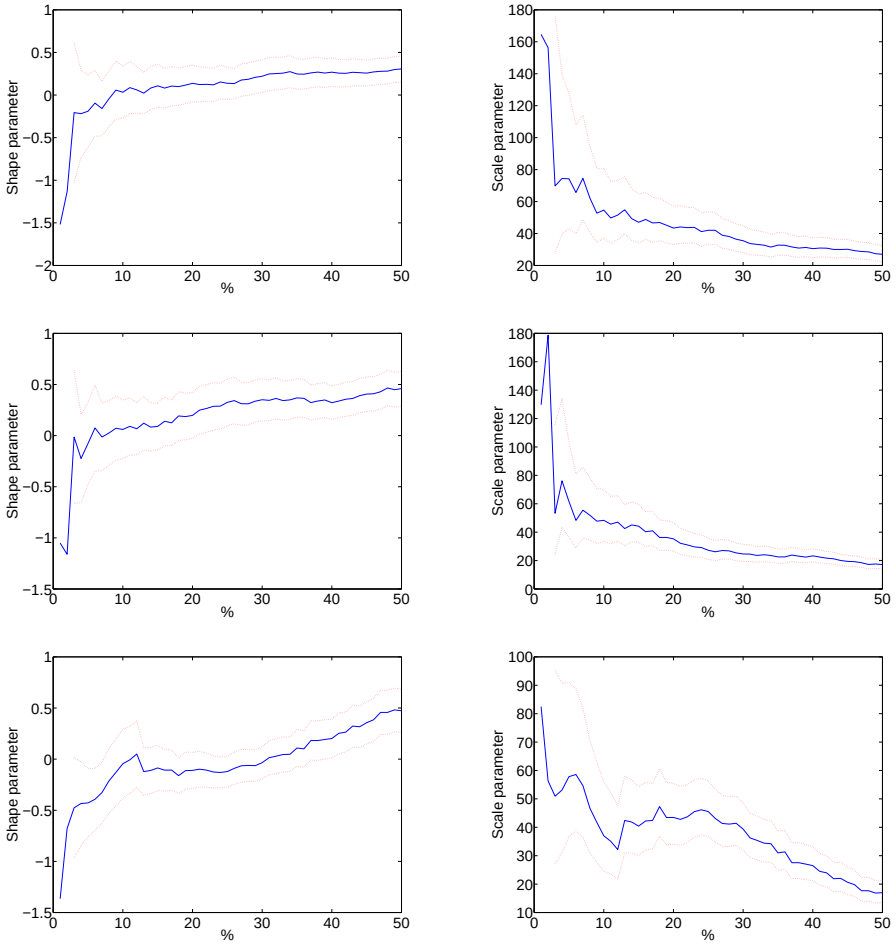

FIG. D.2 – Estimation des paramètre de forme (à gauche) et d'échelle (à droite) par maximum de vraisemblance et intervalles de confiance. En haut : à Barnas, Au milieu : à Mazan-L'Abbaye. En bas : A Valleraugue. Données journalières.

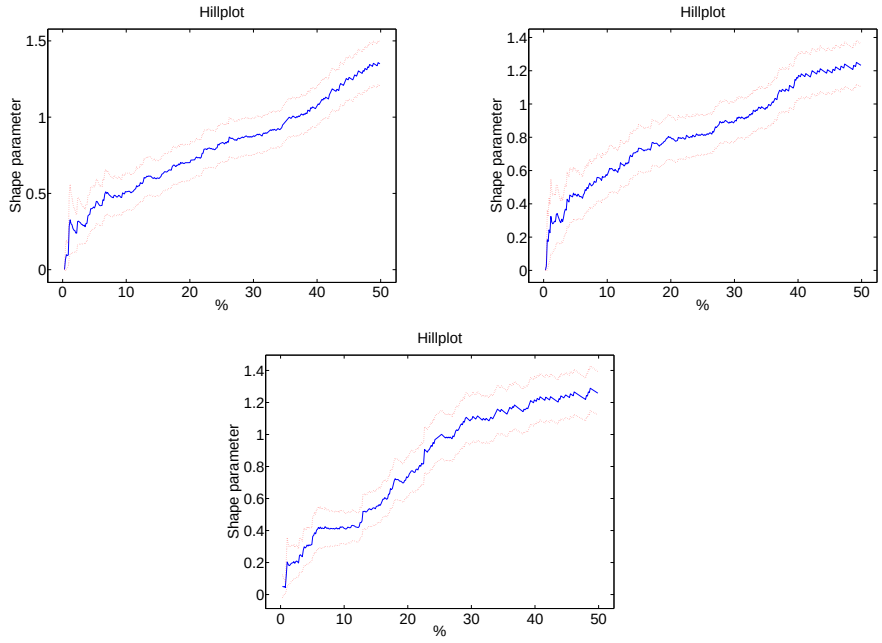

FIG. D.3 – Estimation des paramètre de forme par Hill et intervalles de confiance. En haut : à Barnas, Au milieu : à Mazan-L'Abbaye. En bas : A Valleraugue. Données journalières.

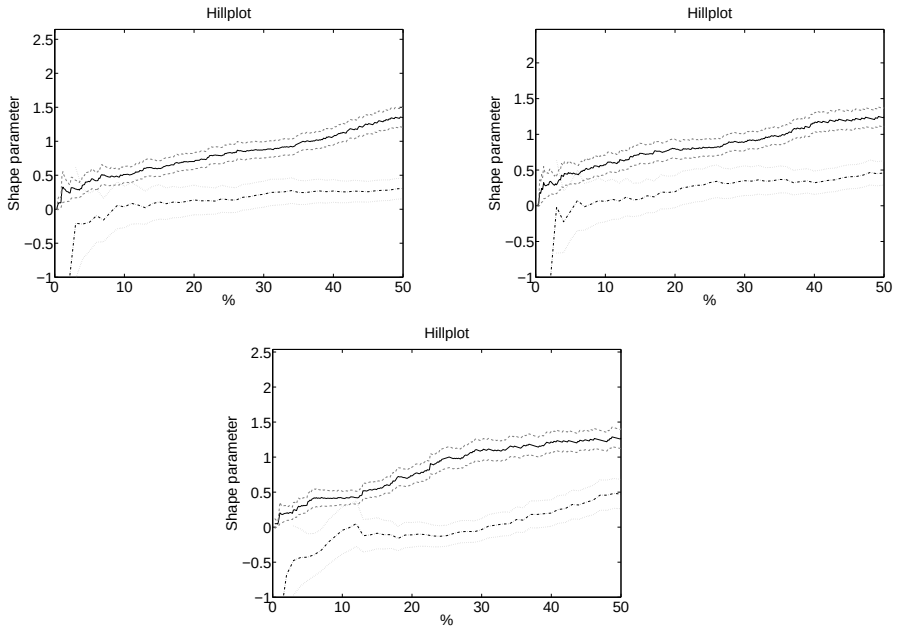

FIG. D.4 – Cohérence des estimations du paramètre de forme par maximum de vraisemblance (trait noir en pointillé) et par Hill (trait noir continu) pour les station de Barnas (en haut à gauche), de Mazan-L'Abbaye (en haut à droite) et de Valleraugue (en bas). Données journalières.

Annexe E

Estimations du paramètre de forme en fonction du seuil

Cette annexe présente l'évolution du paramètre de forme estimé par maximum de vraisemblance et par Hill en fonction du pourcentage d'excès retenus. Les résultats sont présentés pour des données horaires. Avec 5% d'excès, les estimations par maximum de vraisemblance et par Hill diffèrent fortement. En augmentant le pourcentage d'excès, les cartes deviennent similaires avec des valeurs fortes en plaine et faibles en montagne. Cependant les estimations par Hill restent toujours beaucoup plus élevées que par maximum de vraisemblance.

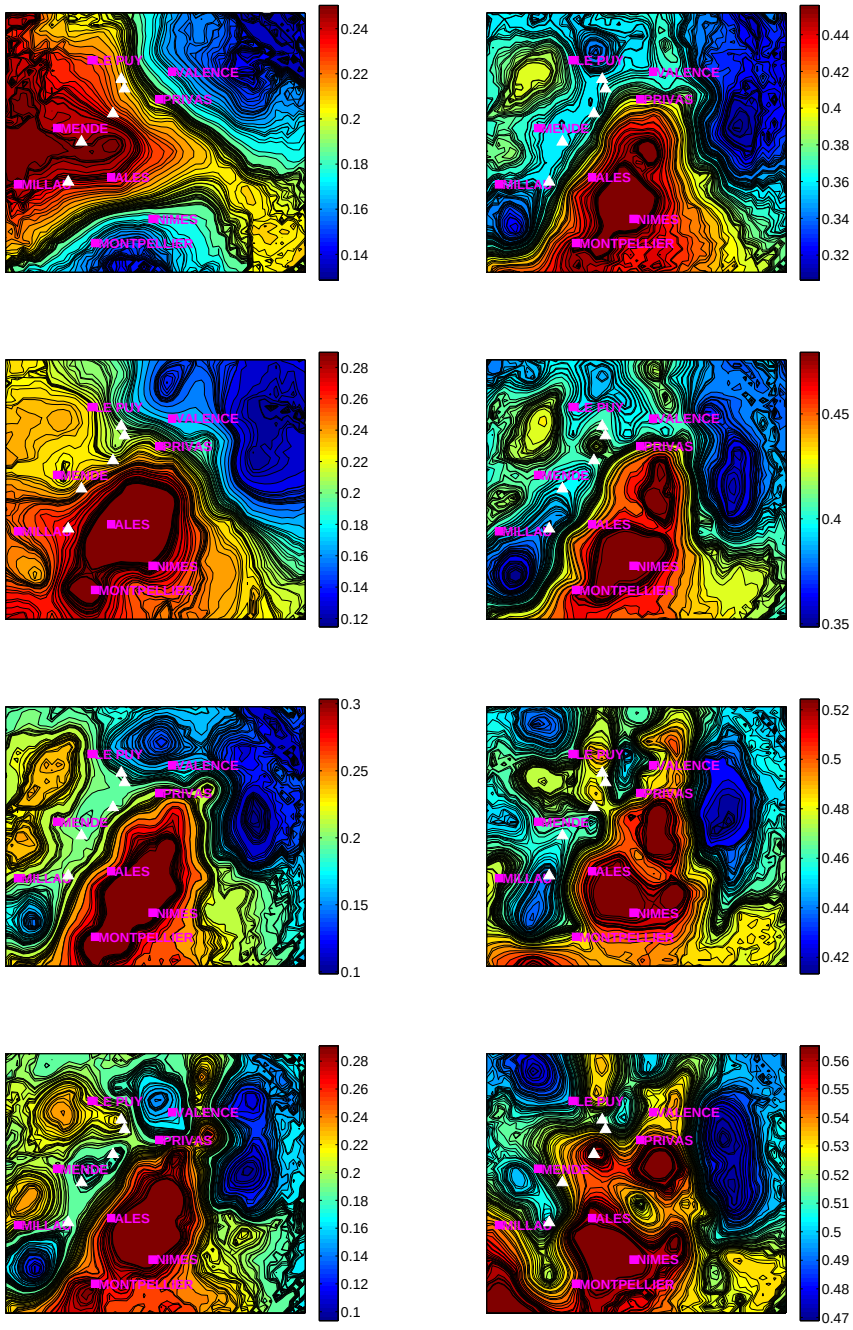

FIG. E.1 – Estimation du paramètre de forme par maximum de vraisemblance (à gauche) et par Hill (à droite) pour un pourcentage d'excès de 5%, 10%, 20% et 30% (de haut en bas)

Annexe F

Variogrammes

Nous présentons dans cette annexe les variogrammes expérimentaux des :
- paramètres de forme
- paramètres d'échelle
- niveaux de retour
- périodes de retour

estimés par maximum de vraisemblance ou par Hill à partir de données horaires ou journalières. Pour chacun des variogrammes, nous présentons le modèle qui a été ajusté et utilisé pour les cartographies par krigeage présentées dans la section 3.

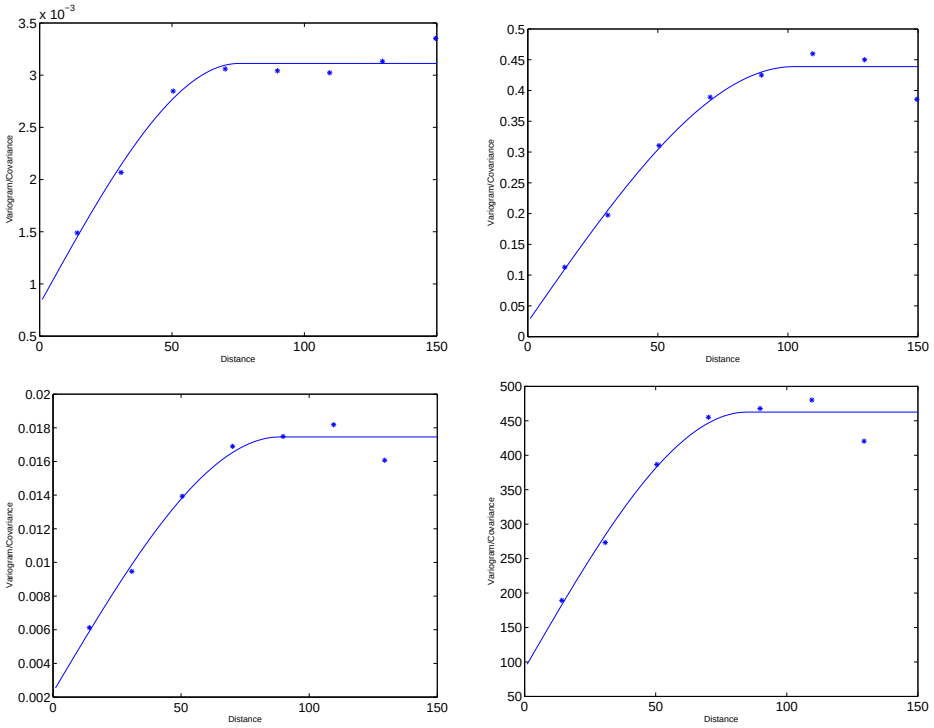

FIG. F.1 – Variogrammes utilisés dans l'étude des données horaires par l'estimateur de Hill. En haut à gauche, variogramme pour le paramètre de forme. En haut à droite : variogramme pour le paramètre d'échelle. En bas à gauche : variogramme pour les niveaux de retour. En bas à droite : variogramme pour les temps de retour.

Données horaires, estimateur de Hill

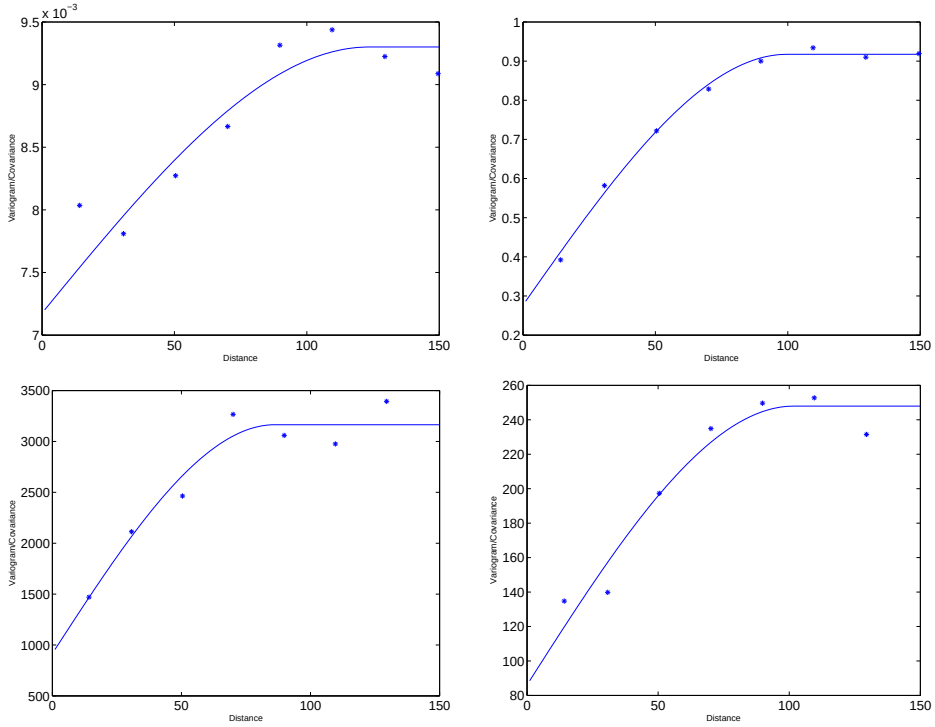

FIG. F.2 – Variogrammes utilisés dans l'étude des données horaires par l'estimateur de maximum de vraisemblance. En haut à gauche, variogramme pour le paramètre de forme. En haut à droite : variogramme pour le paramètre d'échelle. En bas à gauche : variogramme pour les niveaux de retour. En bas à droite : variogramme pour les temps de retour.

Données horaires, maximum de vraisemblance

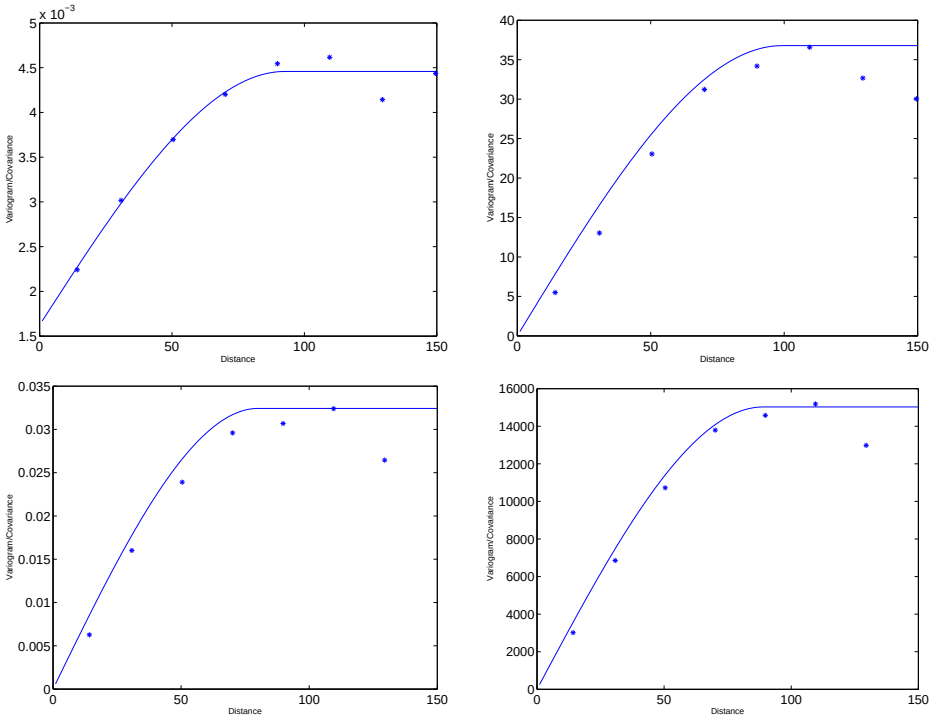

FIG. F.3 – Variogrammes utilisés dans l'étude des données journalières par l'estimateur de Hill. En haut à gauche, variogramme pour le paramètre de forme. En haut à droite : variogramme pour le paramètre d'échelle. En bas à gauche : variogramme pour les niveaux de retour. En bas à droite : variogramme pour les temps de retour.

Données journalières, estimateur de Hill

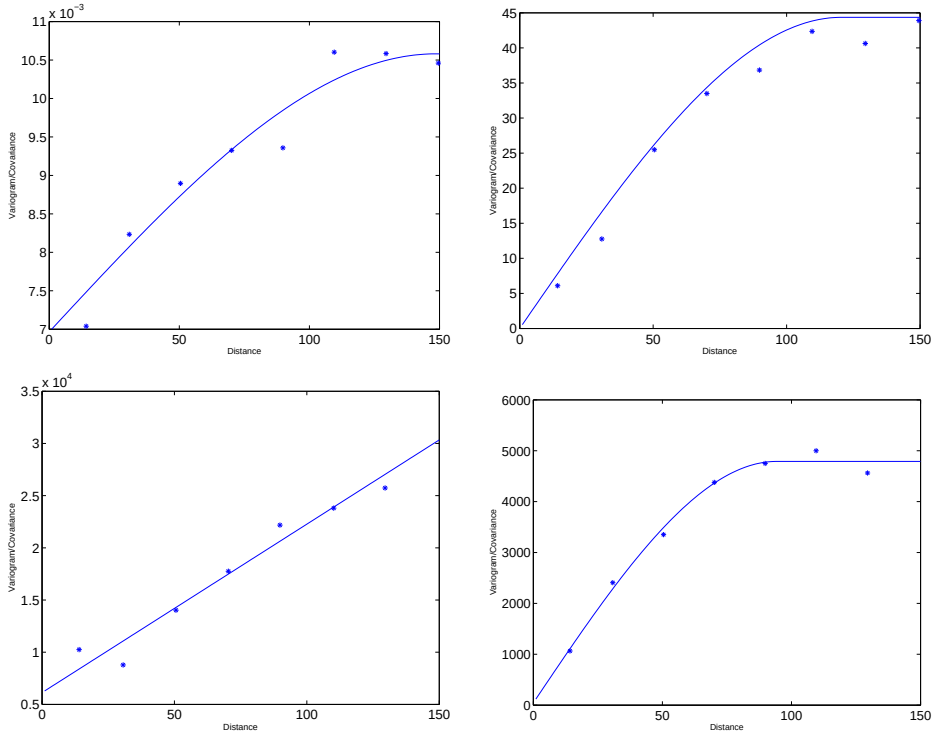

FIG. F.4 – Variogrammes utilisés dans l'étude des données journalières par l'estimateur de maximum de vraisemblance. En haut à gauche, variogramme pour le paramètre de forme. En haut à droite : variogramme pour le paramètre d'échelle. En bas à gauche : variogramme pour les niveaux de retour. En bas à droite : variogramme pour les temps de retour.

Données journalières, maximum de vraisemblance

Bibliographie

[1] L. Bel, J.N. Bacro, and C. Lantuéjoul. Assessing extremal dependance of environmental spatial fields. *Environmetrics*, 18 :1–20, 2007.

[2] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] P. Bois, C. Obled, M.-F. Saintignon, and H. Mailloux. *Atlas expérimental des risques de pluies intenses (2 ème édition)*. Pôle grenoblois detudes et de recherche pour la prévention des risques naturels, 1997.

[4] T.A. Buishand, L. de Haan, and C. Zhou. On spatial extremes : with application to a rainfall problem. *Annals of applied statistics*, 2(2) :624–642, 2008.

[5] V. Chavez-Demoulin and A.C. Davison. Generalized additive models for sample extremes. *Applied Statistics*, 54 :207–222, 2005.

[6] J-P. Chiles and P. Delfiner. *Geostatistics : Modeling Spatial Uncertainty*. Wiley, 1999.

[7] G. Christakos, P. Bogaert, and M.L. Serre. *Temporal GIS : Advanced Functions for Field-Based Applications*. Springer-Verlag, New York, 2002.

[8] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[9] S. Coles and L.R. Pericchi. Anticipating catastrophes through extreme value modelling. *Applied Statistics*, 52(4) :405–416, 2003.

[10] S. Coles, L.R. Pericchi, and S. Sisson. A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, 273 :35–50, 2003.

[11] D. Cooley, P. Naveau, and P. Poncet. Variograms for spatial max-stable random fields. *In Dependence in Probability and Statistics, edited by Bertail P., Doukhan P., Soulier P. ; Springer Lecture Notes in Statistics.*, 187, 2006.

[12] D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association,*, 102(479) :824–840, 2007.

[13] A.C. Davison and R.L. Smith. Models for exceedances over high thresholds. *Journal of the royal statistical society B*, 52(3) :393–442, 1990.

[14] L. de Haan and A. Ferreira. *Extreme value theory : an introduction*. Springer, 2006.

[15] L. de Haan and T.T. Pereira. Spatial extremes : models for the stationary case. *The annals of statistics*, 34(1) :146–168, 2006.

[16] C. Gaetan and M. Grigoletto. A hierarchical model for the analysis of spatial rainfall extremes. *Journal of Agricultural, Biological and Environmental Statistics*, 12(4) :434–449, 2007.

[17] R. Gençay, F. Selçuk, and A. Ulugülyağc. Evim : A software package for extreme value analysis in matlab. *Studies in Nonlinear Dynamics and Econometrics*, 5(3) :213–239, 2001.

[18] M.N. Khaliq, T.B.M.J. Ouarda, J.-C. Ondo, P. Gachon, and B. Bobée. Frequency analysis of a sequence of dependent and/or non-stationary hydrometeorological observations : a review. *Journal of hydrology*, 329 :534–552, 2006.

[19] T. Lebel. *Moyenne spatiale de la pluie sur un bassin versant : estimation optimale, génération stochastique et gradex des valeurs extrmes*. PhD thesis, Thèse de doctorat, Institut National Polytechnique de Grenoble, 1984.

[20] P. Meylan and A. Musy. *Hydrologie fréquentielle*. H*G*A, Bucarest, 1999.

[21] I.D. Morton, J. Bowers, and G. Mould. Estimating return period wave heights and wind speeds using a seasonal point process model. *Coastal Engineering*, 31 :305–326, 1997.

[22] T.P.T Nguyen. *Analyse statistique des valeurs extrêmes de précipitation. Application dans la région Cévennes-Vivarais*. PhD thesis, Thèse de Doctorat, Institut National Polytechnique de Grenoble, 1993.

[23] F. Parisi and R. Lund. Seasonality and return periods of landfalling atlantic basin hurricanes. *Australian and New Zealand Journal of Statistics*, 42(3) :271–282, 2000.

[24] B. Renard, M. Lang, and P. Bois. Statistical analysis of extreme events in a non-stationary context via a bayesian framework : case study with peak-over-threshold data. *Stochastic Environmental Research and Risk Assessment*, 21 :97–112, 2006.

[25] M. Schlather. Models for stationary max-stable random fields. *Extremes*, 5(1) :33–44, 2002.

[26] M. Slimani. *Etude des pluies de fréquence rare faibles pas de temps sur la région Cévennes-Vivarais : estimation, relation avec le relief et cartographie synthétique*. PhD thesis, Thèse de doctorat, Institut de mécanique de Grenoble, Institut National Polytechnique de Grenoble, 1984.

[27] R.L. Smith. Extreme value analysis of environmental time series : an application to trend detection in ground-level ozone. *Statistical Science*, 4(4) :367–393, 1989.

[28] R.L. Smith. Max-stable processes and spatial extremes. *Unpublished manuscript*, 1990.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399